# IMPROVING PATIENT SELECTION FOR SURGICAL TREATMENT OF BONE METASTASES

Olivier Q. Groot

BRITISH MUSEUM

# Improving Patient Selection for Surgical Treatment of Bone Metastases

*Quality of Life, Adverse Events, and the Supplementing Role of Artificial Intelligence*

Olivier Q. Groot

**Improving Patient Selection for Surgical Treatment of Bone Metastases**

PhD thesis, Utrecht University, The Netherlands

© Olivier Q. Groot, Utrecht, 2021

Email: oqgroot@gmail.com

All rights reserved. No part of this thesis may be reproduced or transmitted in any form or by any means without prior written permission from the author. The copyright of the papers that have been published or have been accepted for publication has been transferred to the respective journals.

The research presented in this thesis was financially supported by means of grants for a research fellowship at Massachusetts General Hospital of Harvard Medical School by K.F. Hein Fonds, Annafonds, René Vogels Stichting, Nijbakker-Morra Stichting, KNAW Van Walree, Stichting "De Drie Lichten", Prins Bernhard Cultuurfonds, VSBfonds, Prof. Michaël-van Vloten Fonds, Fundatie van Vrijvrouwe van Renswoude te 's-Gravenhage & Delft, and Vreedefonds. Printing of this thesis was financially supported by Annafonds and Chipsoft.

| | |
|---|---|
| Cover | De Animalium Proprietate, 1565, Manuel Philes |
| | Houghton Library, Harvard University |
| Lay-out | Rens Nitert |
| Printed by | Gildeprint Drukkerijen B.V., Enschede |
| ISBN | 978-94-6419-404-3 |

# Improving Patient Selection for Surgical Treatment of Bone Metastases

Betere selectie van patiënten voor chirurgische behandeling van botmetastasen

(met een samenvatting in het Nederlands)

## Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

donderdag 23 december 2021 des middags te 4.15 uur

door

Olivier Quinten Groot

geboren op 26 januari 1993
te Leiden

| | |
|---|---|
| **Promotoren:** | Prof. dr. J.J. Verlaan |
| | Prof. dr. H.M. Verkooijen |
| **Copromotor:** | Dr. J.H. Schwab |

*Go Pats*

At the start of the pandemic, I happened to be in Europe unable to return 'home' in Boston. Trying to go off the beaten track, I decided to continue my research activities in a vintage camper from the same age I was. My road trip in the Hymer "Hummer" from August 1st until December 15th 2020 took me from Amsterdam to the south of Spain. A PhD on wheels was not what I bargained for, but it turned out to be an attractive alternative for MGH. After all, international connections and collaboration are key to success.

## COLLABORATIVE UNIVERSITIES

- Massachusetts General Hospital, Boston, United States
- John Hopkins, Baltimore, United States
- University of California, Los Angeles, United States
- Memorial Sloan Kettering, New York, United States
- University of Iowa, Iowa City, United States
- University of Vermont, Burlington, United States
- Northeast Ohio Medical University, Rootstown, United States
- National Taiwan University Hospital, Taipei, Taiwan
- Seoul National University College of Medicine, Seoul, South Korea
- Flinders University, Adelaide, Australia
- Hospital Italiano, Buenos Aires, Argentina
- Royal Orthopaedic Hospital, Birmingham, United Kingdom
- Hospital Universitario La Paz, Madrid, Spain
- Instituto Nacional de Cancerología, Bogotá, Colombia
- Rizzoli Orthopaedic Institute, Bologna, Italy
- UMC Utrecht, The Netherlands
- UMC Amsterdam, The Netherlands
- UMC Groningen, The Netherlands
- UMC Leiden, The Netherlands

# TABLE OF CONTENTS

Chapter 1 General Introduction and Thesis Outline 12

PART I. Rising Incidence

Chapter 2 Surgical Treatment of Bone Metastases: National Trends, Clinical Outcomes and Readmission from 2016-2018 32
*Journal of Orthopaedics (Under review)*

PART II. Quality of Life and Physical Function

Chapter 3 Prospective Study for Establishing Minimal Clinically Important Differences in Patients with Surgery for Lower Extremity Metastases 50
*Acta Oncologica 2021*

Chapter 4 Quality of Life Changes after Open Surgery for Spinal Metastases: a Systematic Review and Meta-Analysis. 66
*Clinical Spine Surgery 2021*

Chapter 5 Do Cohabitants Reliably Complete Questionnaires for Patients in a Terminal Cancer Stage when Assessing Quality of Life, Pain, Depression, and Anxiety? 86
*Clinical Orthopaedics and Related Research 2021*

PART III. Mortality and Complications

Chapter 6 Clinical Outcome Differences in the Treatment of Impending Versus Completed Pathological Long Bone Fractures 106
*Journal of Bone and Joint Surgery 2021*

Chapter 7 High Risk of Venous Thromboembolism After Surgery for Long Bone Metastases: A Retrospective Study of 682 Patients 124
*Clinical Orthopaedics and Related Research 2018*

Chapter 8 High Risk of Symptomatic Venous Thromboembolism After Surgery for Spine Metastatic Bone Lesions: A Retrospective Study 142
*Clinical Orthopaedics and Related Research 2019*

Chapter 9 Complications and Reoperations after Surgery for 647 Patients with Spine Metastatic Disease 164
*The Spine Journal 2019*

PART IV. Supplementing Artificial Intelligence Tools

Chapter 10 Body Composition Predictors of Mortality in Patients Undergoing Surgery for Long Bone Metastases 188
*Journal of Surgical Oncology (Under review)*

Chapter 11 Body Composition Predictors of Adverse Postoperative Events in Patients Undergoing Surgery for Long Bone Metastases 206
*Journal of the American Academy of Orthopedic Surgeons (Under review)*

Chapter 12 CT Defined Body Composition Measurements for Prediction of Adverse Postoperative Events in Patients with Spinal Metastases 220
*Spine (Under review)*

Chapter 13 The Preoperative Machine Learning Algorithm for Extremity Metastatic Disease Can Predict 90-day and 1-year Survival: An External Validation Study of 264 Patients 236
*Journal of Surgical Oncology 2021*

Chapter 14 International Validation of the SORG Machine Learning Algorithms for Survival Prediction of Extremity Metastases Undergoing Surgical Treatment 252
*Clinical Orthopaedics and Related Research 2021*

Chapter 15 Comparison of SORG Classical Algorithm and Machine Learning Algorithms for Survival Estimation in Spinal Metastasis: a Meta-Analysis of the Literature 272
*The Spine Journal (Under review)*

Chapter 16 Natural Language Processing for Automated Quantification of Bone Metastases Reported in Free-Text Bone Scintigraphy Reports 292
*Acta Oncologica 2020*

PART V. Strengths and Limitations of Artificial Intelligence

Chapter 17 Machine Learning Prediction Models in Orthopaedic Surgery: a Systematic Review in Transparent Reporting 308
*Journal of Orthopaedic Research 2021*

Chapter 18 Availability and Reporting Quality of External Validations of Machine Learning Prediction Models with Orthopedic Surgical Outcomes: a Systematic Review 324
*Acta Orthopaedica 2021*

Chapter 19 Does Artificial Intelligence Outperform Natural Intelligence in Interpreting Musculoskeletal Radiological Studies? A Systematic Review 344
*Clinical Orthopaedics and Related Research 2020*

PART VI. Summary and General Discussion

Chapter 20 Summary 370

Chapter 21 General Discussion and Future Perspectives 378

APPENDICES

Dutch Summary - Nederlandse samenvatting 396
Authors and Affiliations 402
Review Committee 405
List of Publications 406

CHAPTER

01

# GENERAL INTRODUCTION AND THESIS OUTLINE

## EPIDEMIOLOGY

Bone metastases are the most common form of malignant bone tumors. The earliest evidence of bone metastases are radiographic scans of Egyptian mummies dated from 5000 to 3000 BC.[1] The first medical description waited until 1676 AD, when Richard Wiseman, a barber surgeon in England, described the effects of bone metastases as "rotting the Bones under them" and further, repeatedly emphasized the great suffering of the later stages of the disease (Figure 1).[2]

The seven hallmarks of cancer by Weinberg describe the complexity of cancer. They consist of selective growth, proliferative advantage, resisting cell death, enabling replicative immortality, vascularization, activating invasion, and metastasis.[3] Metastasis occurs when tumor cells spread to a distant location through the blood or lymph system from their primary site where they settle, survive, and grow.[4] Bone is the third most common site of metastasis, after lung and liver.[5,6] Primary tumors of the breast, lung, and prostate are most likely to metastasize to bone.[7] In 2021, 1.9 million new cancer cases, including 400,000 bone metastases, and 600,000 cancer deaths are projected to occur in the United States (population 2021: 331 million).[8] In the Netherlands (population 2019: 17 million), 120.000 were diagnosed with cancer and 46.000 died of cancer in 2019.[9] The incidence of bone metastases has been projected to increase rapidly in the upcoming years due to advances in oncologic care and an ageing population.[10]

The majority of bone metastases are detected incidentally during initial staging, follow-up examination or at treatment reevaluation staging.11 About half of the patients with bone metastases become symptomatic and develop skeletal related events, especially in weight-bearing bones such as femurs and vertebral column, which endure constant dynamic forces.[12, 13] Skeletal related events include bone pain, spinal cord compression, nerve root compression, hypercalcemia, and pathological fracture.[14] These detrimental events cause declined physical function, decreased quality of life, loss of independence and decreased survival. Most patients will present to the hospital to receive multidisciplinary care for their bone metastases in order to prevent and treat skeletal related events and their adverse consequences.[15]

## TREATMENT OF BONE METASTASES

In general, bone metastases are managed by (a combination of) systemic treatment, radiotherapy, and surgery. Every newly diagnosed patient should receive an individualized plan from a multidisciplinary oncological team based on a diagnostic workup including clinical and physical examination, laboratory analysis, radiographic imaging, and preferably, a diagnostic biopsy of the bone lesion for histological confirmation and further receptor/mutation analysis. Various treatment algorithms exist for bone metastases such as the neurologic, oncologic, mechanical, and systemic

**Figure 1**. Radiographic analysis of Egyptian mummies from 5000-3000 BC shows lytic vertebral lesions (left). Richard Wiseman provided the first written description of bone metastases in 1676 AD (right).

(NOMS) decision framework or European Society for Medical Oncology Clinical guidelines, the optimal combination, as well as the sequence of treatments, should have a better understanding. A lot of heterogeneity exists due to numerous variables including patient's wishes and health status, underlying primary tumor, and location of the bone lesion(s).[16,17] Effective use and application of each treatment require a well consolidated multidisciplinary approach and a close collaboration between clinical, radiation, and surgical oncology.

Medical treatment has improved considerably over time with the introduction of systemic treatments such as chemotherapy, hormonal therapy, targeted therapy, and agents to improve bone strength. Although patients are generally treated with chemotherapy and/or hormonal therapy – these are often directed against the primary tumor – the advent of bone targeted agents such as bisphosphonates revolutionized prevention and treatment of skeletal related events by slowing bone loss and strengthening the bone. In selecting a bone targeted agent, the drug, dose, and dosing interval need to be assessed on an individual patient basis including the risk for skeletal related events.[16,18] Over recent years, bone targeted agents have come to be an important adjunct to systemic treatment for bone metastases in patients with breast cancer, multiple myeloma, and other solid tumors.[19]

Palliative radiotherapy is a proven and widely accepted treatment modality particularly for painful bone metastases, except for those lesions considered to be radio resistant.[20,21] Other indications for radiotherapy beside painful lesions include prevention of pathological fractures and neurological complications arising from spinal cord compression, and local tumor control.[22,23] Recent advances such as stereotactic body radiation therapy allow for accurate administration of high doses to metastatic bone with a greater accuracy while sparing adjacent critical structures.[24,25]

With the improvement of the above-mentioned treatment modalities, the need for surgery is much reduced. As a rule of thumb, surgical treatment is prescribed for debilitating (impending) pathological fractures of the spinal column, long bones and hip joints, peripheral nerve compression, or spinal cord involvement.[26] The predominant aims of surgical treatment are pain relief, the preservation of physical function and/or optimizing remaining quality of life. However, invasive strategies can result in serious morbidity due to long periods of hospitalization, discharge to non-home locations, postoperative complications, reoperations, or death from the surgery or subsequent recovery period.[27] Literature reveals that among patients with bone metastases who undergo surgery, the rate of postoperative complications varies from 20% to 47% and reoperations from 10% to 38%.[4,28–32] Survival is generally poor as up to a half of the patients that undergo surgery die within one year.[33–37] Although survival rates have been historically poor, recent medical advances have trended for patients to survive longer. However, prolonged survival introduces a complicating problem with patients returning with new and/or recurrent lesions after index surgery.[10]

Throughout most studies in this thesis, patients with bone metastases who undergo surgery are split up in spinal and long-bone metastases as they often differ in medical urgency, (surgical) treatment regimens, postoperative care and rehabilitation, and complications. For example, in United States studies the median duration of surgery was 3 hours (interquartile range: 2.5-3.5 hours) for long-bone metastases compared with 6 hours (interquartile range: 4.5-7.5 hours) for spinal metastases.[38,39] Despite the differences, both remain a similar patient group with comparable oncological problems, workup, clinical outcomes and, in particular, need for better selection strategies for surgical treatment of bone metastases.

Over the past decade, treatment has evolved from simple decisions regarding the need for either radiotherapy or surgery to multidisciplinary approaches.[16,17] Decision frameworks have been developed to assist in selecting the right patient for the right treatment to optimize patient outcomes. However, current selection strategies can be improved by developing treatment plans for each individual patient. These plans incorporate the likelihood of the aforementioned benefits and risk of possible downsides for each individual patient.

## CHALLENGES IN PATIENT SELECTION FOR SURGICAL TREATMENT

The axiom "the decision is more important than the incision" highlights the importance of personalized harm-benefit analysis that should be applied for each individual patient when considering surgical treatment for bone metastases.[40] Unfortunately, this process of selecting the optimal patient for surgical treatment is far from straightforward, as patients with bone metastases are complex. Patients, together with clinicians, must consider multiple aspects: health status, comorbidities, the underlying primary tumor and its biological behavior in bone, potential adverse surgical events, remaining lifespan, and trade-off choices. One such choice is to prolong life or increase quality of life, the latter of which is often considered to be the most important and difficult to achieve.[12] Thus, patients with bone metastases who are considering surgery are a highly heterogenous population in need for personalized decision tools. In addition, surgeons relying solely on their clinical expertise are notoriously poor at predicting surgical outcome, especially in palliative care.[41] Clinical predictions of survival were performed by three surgeons in 178 patients with incurable abdominal cancer. Prognoses were considered accurate when actual survival fell within a predicted estimation ranging from <1 week to 18-24 months. Prognoses were accurate in 27% and there were substantial differences in predicted survival rates between surgeons. [41,42]

Together with the increasing incidence of bone metastases exacerbating the growing strain on the health care system, a greater understanding of quality-of-life benefits and prevalence, and risk factors of postoperative adverse events is needed when contemplating surgical strategies. Personalized prediction models derived from patient and tumor characteristics may be helpful to improve patient selection and prognostication.[43] However, the majority of these mathematical models are subjected to inaccuracies due to random variation or unknown predictors. A prospective multicenter cohort study of 1.469 patients assessed the clinical accuracy of six commonly cited prognostic scoring tools for patients with spinal metastases. After calculating the score for 1.469 patients, no prognostic scoring system was found to have a good predictive value.[44] Other methods of assessing prognosis should be explored, such as artificial intelligence (AI) prediction models. The potential of AI models in predicting outcomes has been demonstrated by a recent similar study design of 732 patients with spinal metastases. An AI prognostic tool achieved greater accuracy than eight other non-AI models on predicting both 90-day and 1-year survival.[45]

Accurate predictive assessment can guide clinicians and patients in the clinical decision-making progress, and the use of complementing AI tools might be the next step. The landmark study by Lindsey et al. demonstrated that clinicians aided by AI models outperformed clinicians unaided by AI in detecting wrist fractures. On average, clinicians had a relative proportional reduction of misinterpretation when aided by ML models of 47% (95% confidence interval [CI] 37 to 54; $p < 0.001$), compared with their non-aided performance.[46] In bone metastases, few AI tools exist, and the models that are available remain to be validated.

## ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) refers to any technique that enables computers to learn from experience without being explicitly programmed. Algorithms based on AI may have a theoretical advantage due to the ability to implicitly incorporate nonlinear interactions between variables, improve automatically through trial-and-error, and require less formal statistical training.[47] In contrast, common statistical methods like logistic regression require an explicit search for nonlinear relationships and more statistical training to understand modeling techniques and various statistical concepts, including stepwise regression, interactions, and P-values.[48]

There are problems within AI algorithms, however. One major limitation originates from AI's use of unstructured data sources instead of structured data. Structured data are defined and organized in a predefined format most often meticulously constructed by human labor. Unstructured data are a conglomeration of many varied types of data such as text, image, audio and video files. They do not reside in a common database format or spreadsheet file. In medicine, the exponential increase in use of billing and clinical care systems has produced numerous less structured emerging data sources. Use of unstructured datasets creates a bounty of potential problems, from inconsistent and inefficient data availability to patient self-selection unintentionally reinforcing underlying bias.[49] For example, if the data source is filled with stereotypical concepts of gender, the resulting application of AI will extend this bias.[50]

Another problem is that the amount of increased computational power and programming can squeeze out information that does not exist. AI will use whatever inputs are available to achieve the best performance, even exploiting datapoints that may not be reliable.[51] For instance, an AI algorithm was more likely to detect a hip fracture if a radiograph was marked "urgent".[52] Another, non-medical, example is that an AI algorithm did not learn the intrinsic difference between wolves and dogs, but instead classified data based on dogs standing on grass and wolves on snow.[53] Lastly, AI algorithms are commonly referred to as a black box: data goes in, decisions come out, but the operations between input and output are non-transparent.[54] One can look under the hood, but the algorithm is often too complex and decisions untraceable and/or incomprehensible. Nonetheless, the widespread use of electronic health data and increase in computational power has led to unprecedented opportunities for AI algorithms. From predicting postoperative survival to automatically assessing radiographic images, the potential applications of these tools are substantial.

Today, AI tools are widely applied in our daily lives. Email providers can filter spam or propose reply messages, cars recognize traffic signs, and streaming platforms recommend movies based on previous preferences and choices leading to the controversial "rabbit hole". Reinforcement AI algorithms from Google maximize users' engagement by predicting which content would expand their taste rather than feeding existing interests. In other words, one watches many more YouTube videos than the one meeting one's initial interest with the end goal of the algorithm to persuade an individual to spend as much time as possible on the platform.[55]

AI is now entering the realm of orthopaedics at a rapid pace into varying fields from diagnostics to prognostics being a relatively recent development compared to its application in the tech industry, In bone metastases, AI algorithms have also shown great promise in accurately predicting surgical outcomes in individual patients.[45,56,57] Accurate preoperative estimation of 90-day and 1-year survival by AI algorithms – this is based on clinical features including age, comorbidities, primary tumor, visceral metastases, and various preoperative laboratory values – have been reported for both patients with spinal metastases and long-bone metastases undergoing surgical treatment (Figure 2).[45,58] Yet, these algorithms remain to be externally validated, especially in populations with different demographic characteristics compared with the developmental cohort.

**Figure 2.** Scan the QR code to access the freely available AI web-based prediction models on https://sorg.mgh.harvard.edu/

External validation, testing the AI model in a new set of patients not used for development, is necessary to assess the quality and generalizability in different patient populations.[59] Beside validating, critical examination and standardized reporting is warranted to ensure reliable and transparent prediction models. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-Machine Learning (TRIPOD-ML) guideline is currently being developed.[60] In the interim, the non-ML version is proposed to critically assess existing studies.[61] Current prediction algorithms in bone metastases also base their input on clinical data that depend on labor intensive and potentially error-prone manual review of clinical charts. AI algorithms might enable fast and accurate extraction of clinical features from free-text medical notes. Lastly, patients with cancer routinely undergo radiographic images for staging or surveillance. These images might contain "hidden" prognostic parameters such as sarcopenia or decreased visceral fat area that are currently being ignored.[62] AI algorithms can perform automated body composition analyses, which could potentially serve as novel biomarkers of survival or adverse events in patients with bone metastases (Figure 3).

Rather than replacing clinicians in determining surgical necessity, AI can serve as a supplemental tool that may aid both clinician and patient. Before AI algorithms can be considered for routine clinical practice in patients with bone metastases, future studies are needed to validate or refute current predictive models, explore additional prognostic measures, and warrant accurate input data. This ensures clinicians can take full advantage of validated, accurate, and clinically implementable AI decision tools.

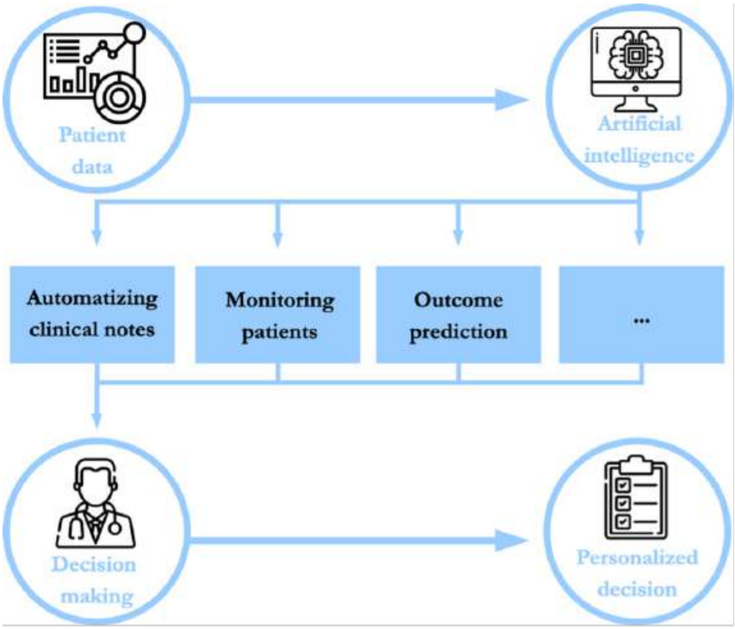


**Figure 3.** Artificial intelligence as the next step towards aiding personalized decision making.

# THESIS OUTLINE

To improve patient selection for surgery of bone metastases, one must first have an accurate understanding of the incidence and consequences of surgery before weighing adverse event risks in comparison to benefits. Therefore, this thesis aims at improving patient selection for surgical treatment of bone metastases by evaluating quality of life outcomes, identifying, and predicting adverse events with the help of AI tools using patient and tumor characteristics, and discussing the challenges associated with these AI tools. **Part I** explores the incidence and outcomes of patients with bone metastases undergoing surgery by using a large, national database representative of the United States. **Part II** studies the quality-of-life benefits of surgical treatment, which are considered the most important outcomes in this vulnerable patient population. **Part III** identifies postoperative adverse events, including mortality, complications, blood transfusions, prolonged hospital stays, and reoperations. These adverse events may substantially undermine the benefits of surgery and have significant impact on patient reported outcomes, primarily quality-of-life benefits. **Part IV** presents AI tools that predict these adverse events and might aid in the decision-making process of choosing the optimal candidate for surgical intervention. **Part V** concludes with a portrayal of the challenges of using AI tools in orthopaedic surgical care based on three reviews (Figure 4).

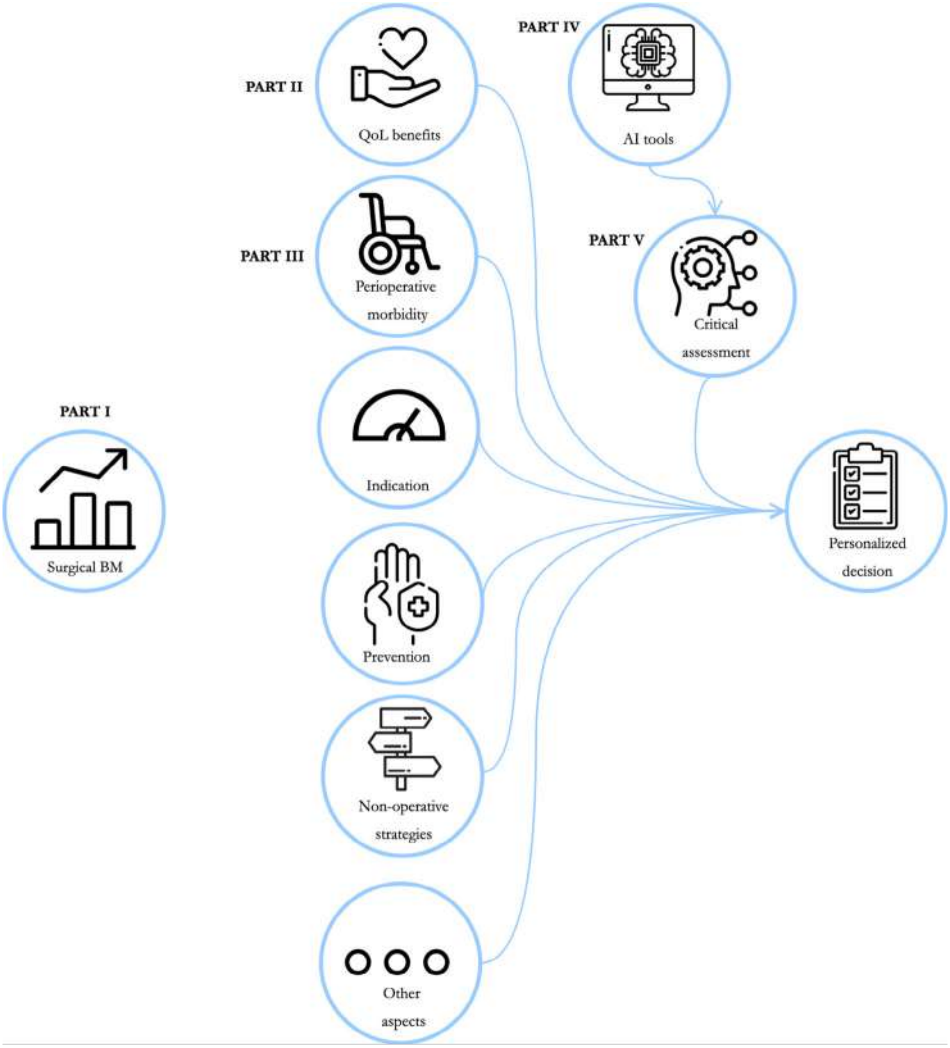


**Figure 4.** Aspects of carefully selecting patients for surgical treatment of bone metastases.

## PART I: RISING INCIDENCE

In recent decades, advances in the treatment of neoplastic disease have prolonged survival for many patients but have resulted – in addition to improved imaging techniques – in increasing frequency of bone metastases.[7] Surgical treatment is commonly offered to patients with bone metastases to improve survival, relieve pain symptoms, and maintain quality of life. Despite the prevalence of bone metastases, overall trends of outcomes, complications, and health-care data on surgical treatment of bone metastases are lacking. **Chapter 2** examines if there is a changing trend in patient demographics, hospital characteristics, complications, and readmissions in patients with bone metastases undergoing surgical treatment.

The Nationwide Readmissions Database dataset was used which contains discharge data from 28 geographically dispersed states, accounting for nearly 60% of all United States hospitalizations. Understanding trends of clinical outcomes and health-care data may guide health-care management, improve the quality of care, and reduce costs in surgical management of bone metastases. In addition, it can help start (non-operative) preventive measures and policies by identifying patients at high risk for detrimental surgical outcomes. Patient selection is critical when considering surgical management, as survival benefits, complication risks, and potential quality of life benefits must all be weighed against one another. Understanding how these trends in mortality and complications are changing over time may help surgeons and patients in management decisions.

## PART II: QUALITY OF LIFE AND PHYSICAL FUNCTION

Quality of life is of increased importance as patients with bone metastases have limited life expectancies, and their disease often causes substantial pain, disability, and psychological stress. This makes assessment of patient reported outcomes vital in understanding and quantifying the effectiveness of surgical treatment on the patients' perceptions about their health. However, the clinical relevance of patient-reported outcomes score changes is often unclear. **Chapter 3** determines this minimal clinically important difference of Patient-Reported Outcomes Measurement Information System (PROMIS) Pain Interference, Cancer-specific Physical Function, and Global (Physical and Mental Health) in patients undergoing surgery for lower extremity metastases. This information is important for patients with bone metastases as surgery is often indicated for palliative purposes. Apart from managing expectations for clinicians and patients during the treatment course, establishing the minimal clinically important difference for quality-of-life questionnaires is expected to aid in the assessment of clinical significance of quality-of-life changes in clinical trials and sample sizes estimates for future studies.

It remains largely unknown to what extent surgery improves quality of life for metastatic spinal disease. Furthermore, studies evaluating quality of life in patients with spinal metastases are using many questionnaires, not all validated for this category of patients, making it difficult to readily compare study results. It would be interesting to quantify the magnitude and duration of quality-of-life benefits after surgery for spinal metastases.[63,64] **Chapter 4** quantifies how surgery affects physical, social/family, emotional, and functional well-being through a systematic review and meta-analysis. These study results can be used to inform patients on postoperative expectations and help physicians to understand the potential postoperative course and use this for decision-making.

Patients with bone metastases often find quality of life questionnaire completion to be physically or emotionally burdensome. Cohabitants (such as spouses, domestic partners, offspring, or other people who live with the patient) could be reliable alternatives to patients as 40% to 70% of patients who are critically ill are unable to complete quality-of-life questionnaires.[65] However, the extent of reliability in this complicated patient population remains undefined, and the influence of the cohabitant's condition on their assessment of the patient's quality of life is unknown. **Chapter 5** investigates whether quality of life scores reported by patients differ markedly from scores as assessed by their cohabitants. These findings can support the use of cohabitants as a reliable alternative to patients who are unable to complete quality of life questionnaires.

## PART III: MORTALITY AND COMPLICATIONS

In addition to quality-of-life, the probability to develop postoperative adverse events including mortality, complications, blood transfusions, prolonged hospital stay, and reoperations has great impact on the decision for/against surgery. These adverse events may substantially undermine the benefits of surgery and have significant impact on patient outcomes, including quality-of-life benefits. A greater understanding of the prevalence and risk factors of these postoperative adverse events can provide valuable insight when contemplating surgical strategies.

The difference in outcome of surgery for an impending versus a completed pathological fracture has not been clearly defined. **Chapter 6** assesses the differences in survival and adverse events between surgical treatment of impending versus completed pathological fractures in long bone metastases. These results may highlight the benefits of prophylactic surgery and emphasize the necessity to accurately predict which bone lesion is at risk to break.

Both cancer and orthopaedic surgery are risk factors for postoperative complications. The risk and prevalence of wound complications for patients undergoing surgery for bone metastases is unknown and it is unclear whether adverse events shorten patients' survival. **Chapters 7 and 8** investigate the risk of venous thromboembolism in both patients with long bone metastases and spinal bone metastases undergoing surgery. A greater understanding of postoperative adverse events is helpful when contemplating surgical treatment. **Chapter 9** assesses the prevalence, types, as well as risk factors for 30-days complications and reoperations in patients with spinal metastases undergoing surgery.

## PART IV: SUPPLEMENTING ARTIFICIAL INTELLIGENCE TOOLS

Artificial Intelligence (AI) algorithms are rapidly emerging tools in medicine, facilitating personalized decision making, diagnostic imaging, and clinical documentation. AI tools can help predict the adverse events to aid the (shared) decision-making process for surgical interventions.

The use of radiographic defined body composition measurements for prognostic purposes is a growing trend in recent oncologic, surgical, and orthopaedic literature. The body composition measurements may serve as imaging biomarkers for predictive purposes including survival, tumor recurrence and complications in patients with and without cancer. **Chapters 10, 11, and 12** investigate the use of automated CT body composition measurements as predictors for mortality and secondary outcomes such as hospitalization, wound complications and reoperations. Especially radiographic measurements from CT are attractive because they are often readily available in the oncologic population and can augment existing prognostication tools.

Accurate preoperative estimation of 90-day and 1-year survival by AI algorithms have been developed for both patients with spinal metastases and long-bone metastases undergoing surgical treatment. Yet, these algorithms remain to be externally validated, especially in populations with different demographic characteristics. In **Chapters 13 and 14**, an existing AI prognostic tool predicting survival in patients undergoing surgery for long bone metastases based on patient and tumor characteristics is validated in both American and Asian cohorts. **Chapter 15** tests the hypothesis that different demographics should be considered by prediction models to ensure accurate and reliable prognoses. Testing of models in data that was not used during development, in particular from populations with different demographic and culturally characteristics, ensures clinicians can take full advantage of validated and clinically implementable AI decision tools.

The widespread availability of electronic health data has led to unprecedented opportunities for automated extraction of clinical features from free-text medical notes. **Chapter 16** investigates if an automated tool accurately extracts from radiology reports meaningful preoperative clinical variables such as number of bone metastases, known to be associated with adverse outcomes in patients with bone metastases. After external validation, these AI algorithms can be integrated into the electronic health care system to supplement procedural or diagnosis codes and bypass error prone and labor-intensive manual chart review to extract meaningful clinical features.

## PART V: STRENGTHS AND LIMITATIONS OF ARTIFICIAL INTELLIGENCE

In the final part of this thesis, three reviews underline the challenges of current AI applications in orthopaedic surgical care. All three chapters include orthopaedic surgery AI studies in general and not in particular bone metastasis. Yet, the data translates to the topic of this thesis as many reviewed studies handle AI algorithms in patients with bone metastases undergoing surgical treatment and non-bone metastases studies provide generalizable messages. **Chapter 17** explores the range of applications and quality of current machine learning prediction models in orthopaedic surgery. This review sheds light, in particular, on transparent reporting of performance measures using the TRIPOD statement, which is necessary to allow accurate evaluation of the machine learning models. It is also imperative that these models are accurate, reliable, and applicable to patients outside the developmental dataset. **Chapter 18** examines the number of available machine learning prediction models that are externally validated. External validation is considered essential before a model can be used in routine clinical practice. Testing the developed model on independent datasets addresses concerns of internal validation, including: the generalizability of the model in different patient populations, shortcomings in statistical modelling (e.g., incorrect handling of missing data), and model overfitting. Lastly, **Chapter 19** investigates where current machine learning developments stand in aiding the clinicians' performance in assessing musculoskeletal abnormalities on imaging. AI models may improve the safety and effectiveness of patient care while working in conjunction with human counterparts rather than replacing clinicians.

# REFERENCES

1. Strouhal E. Ancient Egyptian case of carcinoma. *Bull NY Acad Med.* 1978;54(3):290–302.

2. Wiseman R. Several chirurgical treatises. *London, Royst.* 1676.

3. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* 2000;100(1):57–70.

4. Randall R. Metastatic bone disease: an integrated approach to patient care. *New York Springer-Verlag.* 2016.

5. Wingo PA, Tong T, Bolden S. Cancer statistics, 1995. *CA Cancer J Clin.* 1995;45(1):8–30.

6. Abrams HL, Spiro R, Goldstein N. Metastases in carcinoma; analysis of 1000 autopsied cases. *Cancer.* 1950;3(1):74–85.

7. Viale PH. The American cancer society's facts & figures: 2020 edition. *J Adv Pract Oncol.* 2020;11(2):135–136.

8. Siegel RL, Miller KD, Fuchs HE, et al. Cancer statistics, 2021. *CA Cancer J Clin.* 2021;71(1):7–33.

9. IKC (Integraal Kanker Centrum Nederland). Incidentie en sterfte van kanker. *https://iknl.nl/nkr-cijfers.* 2021.

10. American Cancer Society. Cancer treatment & survivorship facts & figures 2019-2021. *Atlanta Am Cancer Soc.* 2019.

11. Jehn CF, Diel IJ, Overkamp F, et al. Management of metastatic bone disease algorithms for diagnostics and treatment. *Anticancer Res.* 2016;36(6):2631–2637.

12. Coleman RE. Management of bone metastases. *Oncologist.* 2000;5(6):463–470.

13. Oster G, Lamerato L, Glass AG, et al. Natural history of skeletal-related events in patients with breast, lung, or prostate cancer and metastases to bone: a 15-year study in two large US health systems. *Support care cancer.* 2013;21(12):3279–3286.

14. Wilkinson AN, Viola R, Brundage MD. Managing skeletal related events resulting from bone metastases. *BMJ.* 2008;337:a2041.

15. Schulman KL, Kohles J. Economic burden of metastatic bone disease in the U.S. *Cancer.* 2007;109(11):2334–2342.

16. Coleman R, Hadji P, Body JJ, et al. Bone health in cancer: ESMO clinical practice guidelines. *Ann Oncol.* 2020;31(12):1650–1663.

17. Laufer I, Rubin DG, Lis E, et al. The NOMS framework: approach to the treatment of spinal metastatic tumors. *Oncologist.* 2013;18(6):744–751.

18. Barzilai O, Boriani S, Fisher CG, et al. Essential concepts for the management of metastatic spine disease: what the surgeon should know and practice. *Glob spine J.* 2019;9(1 Suppl):98S-107S.

19. So A, Chin J, Fleshner N, et al. Management of skeletal-related events in patients with advanced prostate cancer and bone metastases: incorporating new agents into clinical practice. *Can Urol Assoc J l'Association des Urol du Canada.* 2012;6(6):465–470.

20. Chow R, Hoskin P, Hollenberg D, et al. Efficacy of single fraction conventional radiation therapy for painful uncomplicated bone metastases: a systematic review and meta-analysis. *Ann Palliat Med.* 2017;6(2):125–142.

21. Chow E, Hoskin P, Mitera G, et al. Update of the international consensus on palliative radiotherapy endpoints for future clinical trials in bone metastases. *Int J Radiat Oncol Biol Phys.* 2012;82(5):1730–1737.

22. Saravana-Bawan S, David E, Sahgal A, et al. Palliation of bone metastases-exploring options beyond radiotherapy. *Ann Palliat Med.* 2019;8(2):168–177.

23. De Felice F, Piccioli A, Musio D, et al. The role of radiation therapy in bone metastases management. *Oncotarget*. 2017;8(15):25691–25699.

24. Pielkenrood BJ, van der Velden JM, van der Linden YM, et al. Pain response after stereotactic body radiation therapy versus conventional radiation therapy in patients with bone metastases-a phase 2 randomized controlled trial within a prospective cohort. *Int J Radiat Oncol Biol Phys*. 2021;110(2):358–367.

25. Spencer KL, van der Velden JM, Wong E, et al. Systematic review of the role of stereotactic radiotherapy for bone metastases. *J Natl Cancer Inst*. 2019;111(10):1023–1032.

26. Selvaggi G, Scagliotti G V. Management of bone metastases in cancer: a review. *Crit Rev Oncol Hematol*. 2005;56(3):365–378.

27. Macedo F, Ladeira K, Pinho F, et al. Bone metastases: an overview. *Oncol Rev*. 2017;11(1):321.

28. Versteeg AL, Verlaan JJ, de Baat P, et al. Complications after percutaneous pedicle screw fixation for the treatment of unstable spinal metastases. *Ann Surg Oncol*. 2016;23(7):2343–2349.

29. Jansson KÅ, Bauer HCF. Survival, complications and outcome in 282 patients operated for neurological deficit due to thoracic or lumbar spinal metastases. *Eur Spine J*. 2006;15(2):196–202.

30. Falicov A, Fisher CG, Sparkes J, et al. Impact of surgical intervention on quality of life in patients with spinal metastases. *Spine (Phila Pa 1976)*. 2006;31(24):2849–2856.

31. Shallop B, Starks A, Greenbaum S, et al. Thromboembolism after intramedullary nailing for metastatic bone lesions. *J Bone Jt Surgery-American Vol*. 2015;97(18):1503–1511.

32. Rosen LS, Gordon D, Tchekmedyian NS, et al. Long-term efficacy and safety of zoledronic acid in the treatment of skeletal metastases in patients with nonsmall cell lung carcinoma and other solid tumors: A randomized, phase III, double-blind, placebo-controlled trial. *Cancer*. 2004;100(12):2613–2621.

33. Janssen SJ, Teunis T, Hornicek FJ, et al. Outcome after fixation of metastatic proximal femoral fractures: A systematic review of 40 studies. *J Surg Oncol*. 2016;114(4):507–519.

34. Depreitere B, Ricciardi F, Arts M, et al. How good are the outcomes of instrumented debulking operations for symptomatic spinal metastases and how long do they stand? A subgroup analysis in the global spine tumor study group database. *Acta Neurochir. (Wien)*. 2020;162(4):943–950.

35. Dea N, Versteeg AL, Sahgal A, et al. Metastatic spine disease: should patients with short life expectancy be denied surgical care? An international retrospective cohort study. *Neurosurgery*. 2019.

36. Janssen SJ, van der Heijden AS, van Dijke M, et al. 2015 Marshall Urist Young Investigator Award: prognostication in patients with long bone metastases: does a boosting algorithm improve survival estimates? *Clin Orthop Relat Res*. 2015;473(10):3112–3121.

37. Paulino Pereira NR, Janssen S, Ferrone M, et al. Development of a prognostic survival algorithm for patients with metastatic spine disease. *Spine J*. 2016;16(10):S318.

38. Paulino Pereira NR, Ogink PT, Groot OQ, et al. Complications and reoperations after surgery for 647 patients with spine metastatic disease. *Spine J*. 2019;19(1).

39. Janssen SJ, Kortlever JTP, Ready JE, et al. Complications after surgical management of proximal femoral metastasis: A retrospective study of 417 patients. *J Am Acad Orthop Surg*. 2016;24(7):483–494.

40. Moisi MD, Page J, Gahramanov S, et al. Bullet fragment of the lumbar spine: the decision is more important than the incision. *Glob spine J*. 2015;5(6):523–526.

41. White N, Reid F, Harris A, et al. A systematic review of predictions of survival in palliative care: how accurate are clinicians and who are the experts? *PLoS One.* 2016;11(8):e0161407.

42. Hølmebakk T, Solbakken A, Mala T, et al. Clinical prediction of survival by surgeons for patients with incurable abdominal malignancy. *Eur J Surg Oncol.* 2011;37(7):571–575.

43. Karhade AV, Schwab JH, Del Fiol G, et al. SMART on FHIR in spine: integrating clinical prediction models into electronic health records for precision medicine at the point of care. *Spine J.* 2020.

44. Choi D, Ricciardi F, Arts M, et al. Prediction accuracy of common prognostic scoring systems for metastatic spine disease: results of a prospective international multicentre study of 1469 patients. *Spine (Phila. Pa. 1976).* 2018;43(23):1678–1684.

45. Karhade AV, Thio QCBS, Ogink PT, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. *Neurosurgery*. 2019;85(4);671-681

46. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Nat. Acad Sci USA.* 2018;115(45):11591–11596.

47. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255–260.

48. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol.* 1996;49(11):1225–1231.

49. Chen JH, Asch SM. Machine learning and prediction in medicine - beyond the peak of inflated expectations. *N Engl J Med.* 2017;376(26):2507–2509.

50. Carnevale A, Tangari EA, Iannone A, et al. Will big data and personalized medicine do the gender dimension justice? *AI Soc.* 2021.

51. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):195.

52. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med.* 2019;2:31.

53. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you? Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM; 2016:1135–1144.

54. Medicine TLR. Opening the black box of machine learning. *Lancet Respir Med.* 2018;6(11):801.

55. Roose K. Welcome to the 'Rabbit Hole.' *New York Times (National Ed.)* 2020 Apr

56. Forsberg JA, Eberhardt J, Boland PJ, et al. Estimating survival in patients with operable skeletal metastases: an application of a bayesian belief network. *PLoS One*. 2011;6(5):e19956.

57. Forsberg JA, Wedin R, Boland PJ, et al. Can we estimate short- and intermediate-term survival in patients undergoing surgery for metastatic bone disease? *Clin Orthop Relat Res.* 2017;475(4):1252–1261.

58. Thio QCBS, Karhade AV, Bindels BJJ, et al. Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease. *Clin Orthop Relat Res.* 2020;478(2):322–333.

59. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol.* 2014;14:40.

60. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet (London, England).* 2019;393(10181):1577–1579.

61. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. *BMC Med.* 2018;16(120).

62. Kapoor ND, Twining PK, Groot OQ, et al. Adipose tissue density on CT as a prognostic factor in patients with cancer: a systematic review. *Acta Oncol.* 2020;59(12):1488–1495.

63. Bond MR, Versteeg AL, Sahgal A, et al. Surgical or radiation therapy for the treatment of cervical spine metastases: results from the epidemiology, process, and outcomes of Spine Oncology (EPOSO) Cohort. *Glob Spine J.* 2020;10(1):21–29.

64. Versteeg AL, Sahgal A, Rhines LD, et al. Psychometric evaluation and adaptation of the Spine Oncology Study Group Outcomes Questionnaire to evaluate health-related quality of life in patients with spinal metastases. *Cancer*. 2018;124(8):1828–1838.

65. Jones JM, McPherson CJ, Zimmermann C, et al. Assessing agreement between terminally ill cancer patients' reports of their quality of life and family caregiver and palliative care physician proxy ratings. *J Pain Symptom Manage.* 2011;42(3):354–65.

# PART I

# RISING INCIDENCE

CHAPTER

02

# SURGICAL TREATMENT OF BONE METASTASES: NATIONAL TRENDS, CLINICAL OUTCOMES AND READMISSION FROM 2016-2018

Olivier Q. Groot, Peter K. Twining, Vincent P. Groot, Michelle Shimizu, Neal D. Kapoor, Matthew T. Tobias, Aditya V. Karhade, Helena M. Verkooijen, Jorrit-Jan Verlaan, Joseph H. Schwab

*Journal of Orthopaedics (under review)*

# ABSTRACT

### Background

Advances in the treatment of neoplastic disease have prolonged survival for many patients but have resulted in increasing frequency of patients with bone metastases (BM) undergoing surgery. Understanding trends of clinical outcomes and health-care data may guide health-care management, improve the quality of care, and reduce costs in surgical management of BM.

### Objectives

To examine in-hospital outcomes, 90-day post-discharge readmissions, and trends in operative intervention for bone metastases between 2016-2018 in the United States.

### Design

Epidemiologic study using national administrative data.

### Methods

The Nationwide Readmissions Database (NRD) was used to examine in-hospital outcomes, 90-day post-discharge readmissions, and trends in operative intervention for BM between 2016 and 2018 in the United States. Multivariable logistic regression analyzed outcomes and the Cochran-Armitage method trends.

### Results

The number of surgical BM cases increased from 31,274 in 2016 to 33,361 in 2018, representing a 6.7% increase (P-trend<0.001). Compared with 2016, BM patients in 2018 were older (P=0.015), had more comorbidities (P<0.001), and were more likely to undergo surgery in urban teaching hospitals (P=0.008). Total costs of hospitalization amounted $4.5 billion in 2018. Overall, the incidence of clinical outcomes did not demonstrate a clear trend towards improved or worsened outcomes over all three years.

### Conclusion

The number of BM patients undergoing surgery is increasing, inviting the detrimental outcomes of index hospitalization, and further opening the patient to the increased risk of hospital readmission. Multidisciplinary approaches are needed to formulate individualized plans while also contemplating non-operative strategies in these patients at high-risk of poor outcomes.

# INTRODUCTION

Recent advances in the treatment of neoplastic disease have prolonged survival for many patients, but have resulted in increasing frequency of bone metastases (BM).[1] BM are associated with significant morbidity, mortality, and decreased quality of life.[2] Palliative treatment modalities include chemotherapy, immunotherapy, radiotherapy, and surgery. As a rule of thumb, surgical treatment is commonly indicated in the case of spinal instability or cord compression, and impending or complete pathological fracture through a long-bone metastasis.[3,4] However, it is not without complications as these surgeries are often extensive and invasive, inviting adverse events such as prolonged hospitalization, rehabilitation, postoperative complications, or revision surgery.[1,5]

As a result of the growing number of patients with BM, clinicians and healthcare workers in general are facing an increasing pressure on the healthcare system. Despite the prevalence of BM, overall trends of outcomes, complications, and health-care data on surgical treatment of BM are lacking. These data are needed to reflect modern surgical volume trends, guide health-care management, improve the quality of care, and reduce costs. In addition, it can help start (non-operative) preventive measures and policies by identifying patients at high risk for detrimental surgical outcomes. Patient selection is critical when considering surgical management, as survival benefits, complication risks, and potential quality of life benefits must all be weighed against one another. Understanding how these trends in mortality and complications are changing over time may help surgeons and patients in management decisions.

The purpose of this study, utilizing an USA national database, was to investigate the following: (1) Is there an evolving trend in surgical treatment of BM and what are its effects on clinical outcomes? (2) Do surgical patients with BM have worse clinical outcomes than surgical patients without BM? We hypothesize that there is an increasing incidence of surgical treatment of BM, leading to greater demand for healthcare services. In addition, we hypothesize that patients undergoing surgical treatment for BM have worse outcomes than non-BM patients.

# METHODS

### Study design

This retrospective cohort study utilized data from the Nationwide Readmissions Database (NRD) over a 3-year period from 2016 to 2018.[6] The NRD of 2015 was not considered as it transitioned from using ICD-9 to ICD-10 codes midway through 2015, introducing complexity as trends based on diagnoses or procedures will be affected. The databases prior to 2015 used ICD-9 codes and were therefore disregarded to avoid heterogeneity in the cohort. We also wanted to use the most recent, available years to reflect modern surgical volume trends. The NRD database of 2019 and onwards were not available at the time of analysis.

The NRD data contains discharge data from 28 geographically dispersed states, accounting for nearly 60% of all United States hospitalizations.[6] The database captures all same-state readmissions for a patient within a calendar year, even if the patient presents to a different hospital. It also reports on reasons for hospital visit, associated healthcare costs, length of stay, and discharge location. A stratified algorithm applies weights to each admission, allowing us to project national representative statistics. This study was exempted from the Institutional Review Board review as all data within the database were deidentified and compliant with the Health Insurance Portability and Accountability Act. This study adhered to the guideline of Strengthening the Reporting of Observational studies in Epidemiology (STROBE).[7]

### Patient Selection

All patients with ICD-10 procedure codes that confirmed surgery to the bone, including resection, excision, insertion, replacement, resection, exploration, stabilization, and decompression or release of the spinal cord were included, irrespective of indication (Appendix 1). Patients younger than 18 years were excluded. Patients were then separated into the "BM" or "non-BM" cohort depending on whether they had an ICD-10 diagnosis code for BM (Figure 1). These selection criteria were applied to the first unique hospital admission of each database year – "the index hospitalization".

### Outcomes and Explanatory Variables

The following explanatory variables were considered: age, sex, insurance payer type (Medicare, Medicaid, private, self, other), median household income quartile based on patient's zip code, emergent surgery, hospital size (small, medium, large), hospital type (rural, urban nonteaching, urban teaching), risk of mortality and loss of function (minor, moderate, major, extreme) using the All Patient Refined Diagnosis Related Groups (APR-DRG), which classifies patients based on clinical similarities and their use of hospital resources,[8] and the Elixhauser Comorbidity Index based on a previously developed ICD-10 coding algorithm.[9]

The primary outcomes during index hospitalization included mortality, blood transfusion, postoperative complications using verified complication codes,[10] length of stay (LOS), hospital costs, and discharge location. After adjusting for inflation between 2016 and 2018 with the USA Consumer Price Index, an increased charge of costs was defined as any value over the median cost ($86,893) of a surgical procedure within the BM group.[11] Extended LOS was defined using the same approach as any stay longer than 7 days. The secondary outcomes included overall hospital readmission, readmission for infection,[10] readmission for embolism, and readmission for revision surgery, all

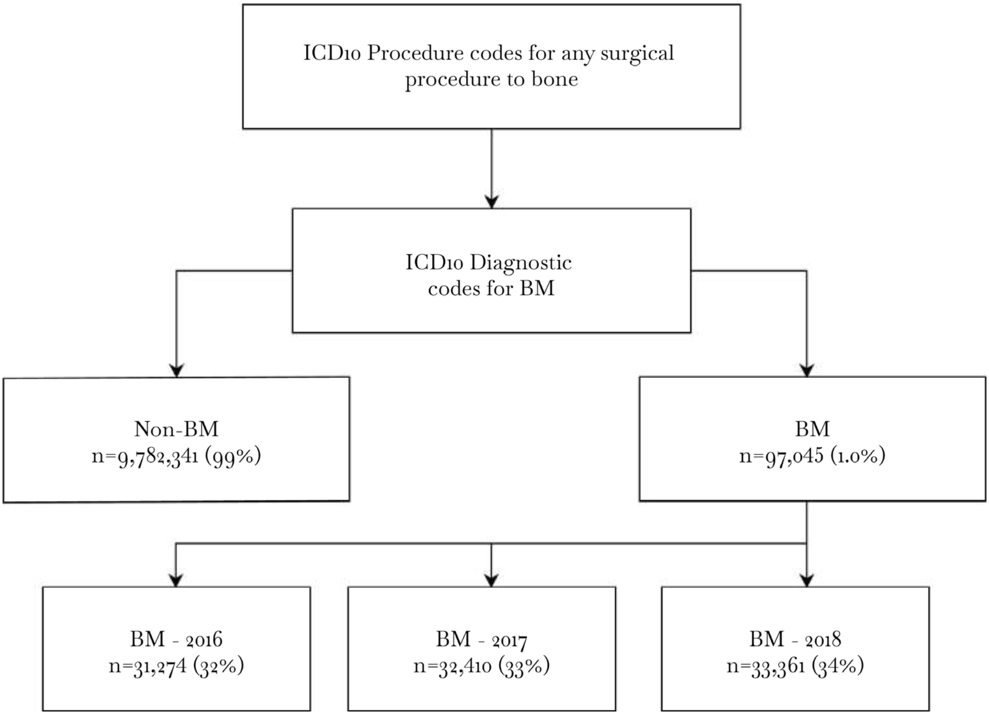


**Figure 1.** Flowchart of selection criteria and included patients.

within 90 days post-discharge of index surgery (Appendix 1).

### Statistical Analysis

All results represent national representative statistics by using the NRD's two-stage stratified algorithm and the Stata's survey (svy) command. Baseline characteristics were compared between the two groups by using t-test for continuous variables – as continuous variables were normally distributed – and Chi squared test for categorical variables. The Cochran-Armitage method tested for a linear trend in proportions by assuming that the proportions follow a linear trend on the logistic scale.[12] BM patients were grouped separately into 2016 and 2018 cohorts to detect any changes over time in both baseline characteristics and outcomes. A multivariable logistic regression assessed the effect of BM on primary and secondary outcomes by controlling for age, sex, Elixhauser Comorbidity Index, emergent surgery, insurance, income, hospital size, and hospital type. Odd ratios (OR) with 95% confidence intervals (CI) were provided. In addition, multivariable logistic regression was used to assess risk factors for both primary and secondary outcomes within the BM group while correcting for all explanatory variables with a P-value <0.10 from univariate testing, Patients with missing outcomes were excluded from analyses as the highest missing value among all outcomes was less than 1.0%: <0.1% (263/9,879,386) for LOS and 0.8% (80,521/9,879,386) for total costs. No missing data was present for the explanatory variables. Patients who died during admission were excluded from calculating readmission rates. A two-tailed P-value of <.05 was considered significant. All statistical analyses were performed using Stata 15.0 (StataCorp LP, College Station, TX, USA).

# RESULTS

A total of 9,879,386 (weighted) patients underwent surgery to the bone from 2016 to 2018 in the United States. Among these patients, 9,782,341 (99%) had a non-BM diagnosis and 97,045 (1.0%) had a BM diagnosis (Figure 1).

The number of surgical cases for BM increased from 31,274 in 2016 to 33,361 in 2018, representing a 6.7% increase (p-trend<0.001; Figure 2). Compared with BM patients in 2016, those in 2018 were older (p=0.015), had a higher Elixhauser Comorbidity Index (p<0.001), and were more likely to undergo surgery in urban teaching hospitals (p=0.008; Table 1). BM patients undergoing surgery in 2018 had an increased rate of postoperative neurological complications (OR, 1.46; 95%CI, 1.31–1.64; p<0.001), other complications (OR, 1.35; 95%CI, 1.26–1.45; p<0.001), non-home discharge location (OR, 1.08; 95%CI, 1.01–1.15; p=0.024), 90-day readmission for infections (OR, 1.12; 95%CI, 1.04–1.20; p=0.003), a decreased rate of mortality (OR, 0.84; 95%CI, 0.73–0.96; p=0.010), and postoperative respiratory complications (OR, 0.90; 95%CI, 0.84–0.95; p=0.001; Table 2). Overall, the incidence of both primary

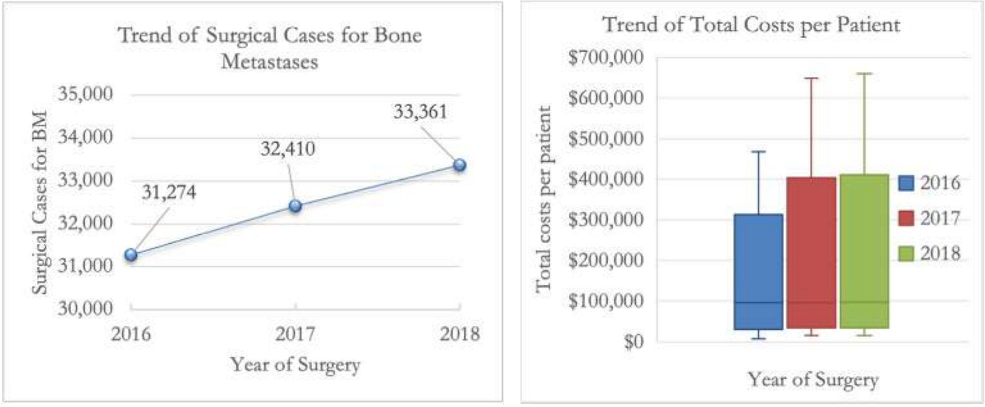


**Figure 2.** (left) The number of surgical cases for BM increased from 31,274 in 2016 to 33,361 in 2018, representing a 6.7% increase (p-trend<0.001). (right) The median total costs per patient per primary hospitalization with interquartile range was $95,801 (53,026-158,391) for 2016; $95,174 (53,883-159,872) for 2017; and $96,615 (54,886-162,474) for 2018.

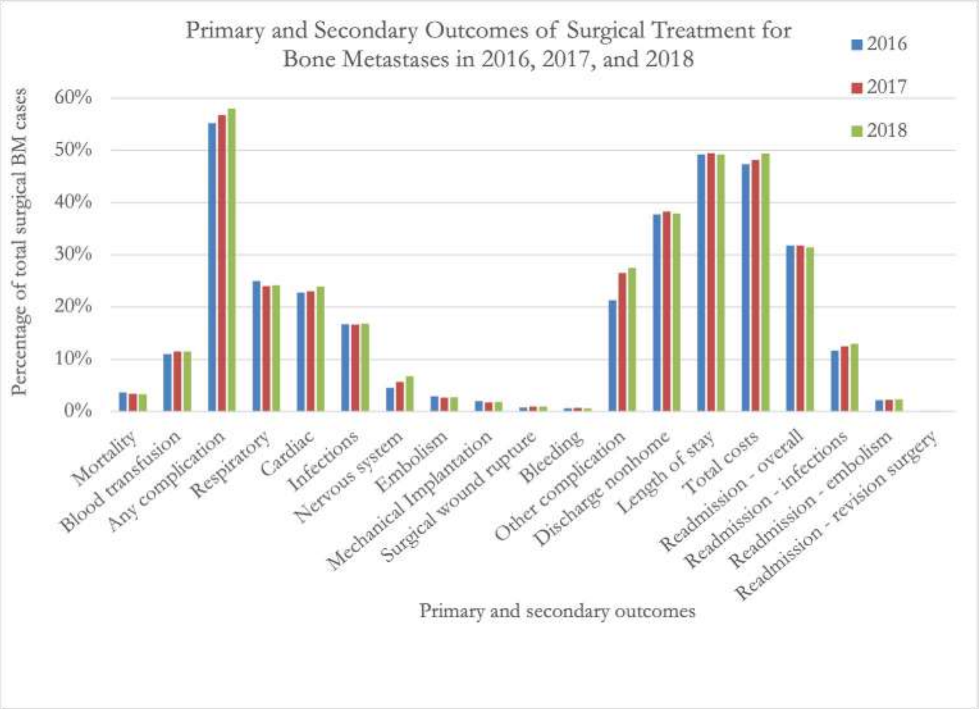


**Figure 3.** The incidence of primary and secondary outcomes in surgical cases for BM per year. Overall, no clear trend was observed towards improved or worsened outcomes.

and secondary outcomes did not demonstrate a clear trend towards improved or worsened outcomes over all three years (Figure 3). The total costs for the primary surgical treatment and hospitalization amounted $4.5 billion in 2018. Independent risk factors for all outcomes are listed in Appendix 2.

Compared with non-BM patients, BM patients were older, more often male, had a higher Elixhauser Comorbidity Index, underwent more emergent surgery on weekends in large, urban teaching hospitals, and had Medicare as insurance (all p<0.001; Appendix 3). During index hospitalization, BM patients had an increased risk of death (OR, 3.02; 95% CI, 2.84–3.21; p<0.001), blood transfusion (OR, 1.63; 95% CI, 1.55–1.72; p<0.001), any postoperative complication (OR, 1.17; 95% CI, 1.14–1.20; p<0.001), non-home discharge location (OR, 1.09; 95% CI, 1.06–1.12; p<0.001), LOS longer than 7 days (OR, 4.07; 95% CI, 3.90–4.23; p<0.001), and total cost of more than $86,893 (OR, 2.22; 95% CI, 2.13–2.32; p<0.001) as compared with patients without BM. BM patients had an increased risk of overall readmission (OR, 2.64; 95% CI, 2.57–2.71; p<0.001), readmission for infection (OR, 2.44; 95% CI, 2.36–2.52; p<0.001), readmission for embolism (OR, 4.05; 95% CI, 3.79–4.33; p<0.001), and a decreased risk of revision surgery (OR, 0.70; 95% CI, 0.52–0.95; p=0.021; Table 3) within 90 days of surgery.

**Table 1.** Demographics and hospital characteristics of surgical BM patients between 2016 and 2018.

| Variables | 2016<br>n=31,274 | 2018<br>n=33,361 | P-value |
|---|---|---|---|
| | mean (95% CI) | | |
| Age (years) | 65.6 (65.3-66.0) | 66.3 (66.0-66.7) | **0.015** |
| Elixhauser Comorbidity Index | 3.4 (3.3-3.4) | 3.5 (3.5-3.6) | **<0.001** |
| | % | % | |
| Female | 48% | 47% | 0.539 |
| Risk of mortality | | | **<0.001** |
| Minor | 1.8% | 1.4% | |
| Moderate | 45% | 42% | |
| Major | 42% | 45% | |
| Extreme | 11% | 12% | |
| Loss of function | | | **<0.001** |
| Minor | 5.4% | 4.2% | |
| Moderate | 34% | 30% | |
| Major | 49% | 49% | |
| Extreme | 12% | 16% | |
| Emergent surgery | 71% | 73% | 0.241 |
| Admission in weekend | 16% | 16% | 0.142 |
| Insurance | | | 0.347 |
| Medicare | 57% | 58% | |
| Medicaid | 11% | 10% | |
| Private | 28% | 27% | |
| Self | 1.6% | 1.8% | |
| Other | 2.7% | 2.9% | |
| Income (percentile) | | | 0.837 |
| 0-25th | 25% | 25% | |
| 25-50th | 27% | 27% | |
| 50-75th | 26% | 26% | |
| 75-100th | 22% | 22% | |
| Hospital size | | | 0.881 |
| Small | 11% | 11% | |
| Medium | 21% | 22% | |
| Large | 68% | 67% | |
| Hospital type | | | **0.008** |
| Urban teaching | 79% | 83% | |
| Urban non-teaching | 17% | 14% | |
| Rural | 3.6% | 3.4% | |

*BM=bone metastases; CI=confidence intervals. All P-values are calculated with the Chi-squared or two-tailed Student's t test.* **Bold** *indicates significance (P<0.05).*

Table 2. Primary and secondary outcomes for surgical BM patients between 2016 and 2018.

| Outcomes | 2016 % (n)<br>n=31,274 | 2018 % (n)<br>n=33,361 | OR<br>(95% CI) | SE | P-value |
|---|---|---|---|---|---|
| *Index hospitalization* | | | | | |
| Mortality | 3.7% (1,148) | 3.3% (1,114) | 0.84 (0.73-0.96) | 0.058 | **0.010** |
| Blood transfusion | 11% (3,446) | 11% (3,819) | 0.99 (0.84-1.19) | 0.090 | 0.995 |
| Postoperative complications | | | | | |
| Any | 55% (17,263) | 58% (19,356) | 1.03 (0.97-1.10) | 0.034 | 0.366 |
| Respiratory | 25% (7,800) | 24% (8,050) | 0.90 (0.84-0.95) | 0.028 | **0.001** |
| Cardiac | 23% (7,124) | 24% (7,986) | 1.07 (0.99-1.14) | 0.037 | 0.062 |
| Infections | 17% (5,226) | 17% (5,614) | 0.95 (0.88-1.02) | 0.036 | 0.172 |
| Nervous system | 4.5% (1,414) | 6.7% (2,248) | 1.46 (1.31-1.64) | 0.084 | **<0.001** |
| Embolism | 2.9% (907) | 2.7% (907) | 0.88 (0.77-1.01) | 0.061 | 0.070 |
| Mechanical implantation | 2.0% (632) | 1.9% (623) | 0.91 (0.76-1.10) | 0.087 | 0.338 |
| Surgical wound dehiscence | 0.8% (244) | 0.9% (303) | 1.16 (0.87-1.53) | 0.166 | 0.311 |
| Bleeding | 0.6% (191) | 0.6% (210) | 1.03 (0.77-1.37) | 0.149 | 0.851 |
| Other | 21% (6,652) | 28% (9,177) | 1.35 (1.26-1.45) | 0.047 | **<0.001** |
| Discharge location non-home | 38% (11,809) | 38% (12,630) | 1.08 (1.01-1.15) | 0.035 | **0.024** |
| Length of stay longer than 7 days | 49% (15,421) | 49% (16,430) | 0.94 (0.86-1.02) | 0.039 | 0.112 |
| Total costs | | | | | |
| More than $86.893 per patient* | 47% (14,808) | 49% (16476) | 1.08 (0.95-1.22) | 0.071 | 0.268 |
| In billions per year ($) | 4.1 | 4.5 | - | - | 0.059 |
| *Hospital readmission within 90 days of surgery* | | | | | |
| Overall | 32% (9,926) | 31% (10,478) | 0.98 (0.93-1.03) | 0.027 | 0.424 |
| Infections | 12% (3,643) | 13% (4,330) | 1.12 (1.04-1.20) | 0.041 | **0.003** |
| Embolism | 2.2% (682) | 2.3% (767) | 1.05 (0.90-1.23) | 0.083 | 0.518 |
| Revision surgery | 0.2% (47) | 0.1% (47) | 0.88 (0.42-1.85) | 0.334 | 0.735 |

*BM=bone metastases; OR=odds ratios; CI=confidence interval; SE=standard error; IQR=interquartile range. All P-values calculated by multivariate logistic regression after correcting for age, Elixhauser comorbidity, gender, emergent surgery, admission in weekend, type of insurance, income, hospital size, and hospital type.* **Bold** *indicates significance (P<0.05).*

**The median with interquartile range is $95,801 (53,026-158,391) for 2016 and $96,615 (54,886-162,474) for 2018.*

**Table 3.** Primary and secondary outcomes in surgical patients with non-BM and BM.

| Outcomes | Non-BM %<br>n=9,782,341 | BM %<br>n=97,045 | OR<br>(95% CI) | SE | P-value |
|---|---|---|---|---|---|
| *Index hospitalization* | | | | | |
| Mortality | 0.6% (53,803) | 3.5% (3,357) | 3.02 (2.84-3.21) | 0.093 | <0.001 |
| Blood transfusion | 4.5% (442,161) | 11% (10,995) | 1.63 (1.55-1.72) | 0.042 | <0.001 |
| Postoperative complications | | | | | |
| Any | 40% (3,885,545) | 57% (55,014) | 1.17 (1.14-1.20) | 0.016 | <0.001 |
| Respiratory | 14% (1,340,180) | 24% (23,620) | 1.61 (1.57-1.66) | 0.024 | <0.001 |
| Cardiac | 19% (1,897,774) | 23% (22,562) | 1.41 (1.37-1.45) | 0.020 | <0.001 |
| Infections | 12% (1,129,860) | 17% (16,235) | 1.28 (1.24-1.32) | 0.021 | <0.001 |
| Nervous system | 2.9% (284,666) | 5.7% (5,512) | 1.08 (1.02-1.13) | 0.027 | 0.004 |
| Embolism | 0.4% (43,042) | 2.8% (2,678) | 3.85 (3.62-4.11) | 0.124 | <0.001 |
| Mechanical implantation | 3.5% (346,294) | 1.9% (1,814) | 0.48 (0.45-0.52) | 0.019 | <0.001 |
| Surgical wound rupture | 0.4% (40,107) | 0.9% (853) | 1.34 (1.20-1.49) | 0.075 | <0.001 |
| Bleeding | 0.3% (26,412) | 0.6% (611) | 1.30 (1.15-1.48) | 0.083 | <0.001 |
| Other | 13% (1,259,965) | 25% (24436) | 1.25 (1.22-1.29) | 0.018 | <0.001 |
| Discharge location non-home | 23% (2,214,722) | 38% (36,858) | 1.09 (1.06-1.12) | 0.017 | <0.001 |
| Length of stay longer than 7 days[a] | 12% (1,213,988) | 49% (47,891) | 4.07 (3.90-4.23) | 0.084 | <0.001 |
| Total costs more than $86.893[b] | 24% (2,341,892) | 48% (46,911) | 2.22 (2.13-2.32) | 0.050 | <0.001 |
| *Hospital readmission within 90 days of surgery* | | | | | |
| Overall | 11% (1,030,080) | 32% (30,705) | 2.64 (2.57-2.71) | 0.036 | <0.001 |
| Infections | 3.4% (332,599) | 12% (12,004) | 2.44 (2.36-2.52) | 0.041 | <0.001 |
| Embolism | 0.4% (38,151) | 2.2% (2,164) | 4.05 (3.79-4.33) | 0.139 | <0.001 |
| Revision surgery | 0.1% (13,695) | 0.1% (135) | 0.70 (0.52-0.95) | 0.108 | 0.021 |

*BM=bone metastases; OR=odds ratios; CI=confidence interval; SE=standard error; IQR=interquartile range. All P-values are calculated by multivariable logistic regression after correcting for age, Elixhauser comorbidity, gender, emergent surgery, admission in weekend, type of insurance, income, hospital size, and hospital type.*
*Missing outcomes included <0.1% (263) for length of stay and 0.8% (80,521) for total costs.*
*a The mean length of stay was 3 (IQR 2-4) for non-BM and 6 (IQR 4-11) for BM.*
*b The median total costs was $56,795 (IQR $33,768-90,211) for non-BM and $93,638 (IQR $52,801-156,856) for BM.*

# DISCUSSION

An increasing number of patients with BM are undergoing surgery. Our study demonstrates that any surgery to the bone is not without risk, especially for BM that are associated with increased risk of death, complications, length of hospital stay, non-home discharge, cost, and readmission as compared with non-BM surgeries. No clear trend was observed regarding improved outcomes over the years 2016-2018 as some outcomes improved while others declined. To our knowledge, this is the first study examining the trends, clinical outcomes, and utilization of healthcare resources in surgical care of BM using an USA national representative database. These findings stress the increased pressure on the healthcare system by patients with BM who are at high-risk for poor outcomes and highlight the importance of preparation and anticipation. For example, preventive measures and policies should facilitate other non-operative treatment strategies such as early-on radiotherapy for patients at high-risk for surgical detrimental outcomes.

This study has several limitations. First, a national database using codes is not as reliable as manual review of medical records or prospective studies. However, these extremely large data sets allowed us to highlight general trends over time. The accuracy of national databases can be improved by using artificial intelligence aided tools such as natural language processing (NLP), which have proven be more accurate in reporting of clinical outcomes than ICD codes.[13] This NLP tool can also be used to determine demographical and clinical characteristics.[14] Second, only same-state readmissions were included. In general, surgical treatment of BM is complex and patients may have to travel across state borders to specialized institutions. However, a readmission for a superficial wound infection or venous thromboembolism would most likely occur in their local state hospital. Therefore, our 90-day readmission rate is likely to be an underestimation. Third, our study does not assess changes in quality of life after surgery, which is arguably the most important factor to consider in patient with advanced metastatic disease. Fourth, the surgery indications could not be determined which would have been interesting to investigate for any trends. Not knowing the indication also meant that patients with BM could have been operated on a non-BM related problem while being grouped in the BM cohort based on the ICD-10 code for BM. For example, a patient with spinal metastases falls and undergoes surgery for a non-cancer related hip fracture. We anticipate a relatively low number of these patients with minor influence on our results. Fifth, only the years 2016 to 2018 were included in this study. While we were able to show several trends over this time period, 3 years may not be a sufficient time period to assess how changing trends are affecting patient care, as advances in medicine often take many years to become fully implemented.[15] These years were chosen because prior to 2016, the NRD dataset used ICD-9 codes, and data for 2019 and onward was not available. Despite having only three years of data, we thought it valuable to provide an up-to-date display of recent trends in surgical treatment of BM.

Finally, risk factors for outcomes listed in Appendix 2 are not as definite as many other known confounding factors are not included in the multivariate analysis. For example, visceral metastases and laboratory values are known factors associated with embolism or mortality.[16,17] However, the identified risk factors in this study may still be useful as they indicate on a population level which factors are important, including income level, hospital size, admission on the weekends, and emergency surgery. For outcomes prediction, we recommend using identified risk factors in subpopulations, ideally using personalized predictions models such as NEMS, SORG, or PATHFx.[16,18,19] Despite these limitations, this study shows several trends in the surgical management of BM which underscore the increasing burden these patients will place on the healthcare system.

The number of patients surgically treated for BM increased from 2016 to 2018, which is consistent with two previous studies on the trends in treatment of spinal metastases from 2000 to 2009 and from 2010 to 2014.[5,20] The survival rates of cancers have been increasing in recent years due to advances in treatments. The improved survival prognoses of cancer may result in increased metastases to bone, which is reflected in the increased cases of surgical treatment of BM, placing a burden on both musculoskeletal surgeons and oncological care in general.[21] To optimally treat patients with BM, the musculoskeletal community needs to understand the incidence, patient demographics, and clinical outcomes.

The rapid pace of advances in oncologic care is reflected in the changing demographics of patients with BM from 2016 to 2018: patients in 2018 were older, had more comorbidities, a higher pre-operative risk for mortality and post-operative complications, and were increasingly operated on at urban teaching hospitals. Clinicians are facing an increasingly complex patient group, who are seeking care at urban teaching hospitals. As urban teaching hospitals transform into more highly specialized centra in order to provide adequate care to these complicated patients, our health care system becomes increasingly costly. This increased price tag is reflected in the growing trend of total costs per patient.

Despite the increased complexity and pre-operative risk of these patients, inpatient mortality decreased slightly from 3.7% in 2016 to 3.3% in 2018. This decreased mortality may reflect the advances in systemic therapy, improved postoperative care, or the implementation of new surgical techniques. However, the trends in post-operative complications and other adverse outcomes have not declined as one would have expected with the improvement in postoperative management. One reason could be that patients in 2018 underwent surgery with, as detailed above, unfavorable demographics and comorbidities which outweighed the improvement in healthcare. As surgical BM cases increase, clinicians need to remain aware that while mortality associated with surgery for BM is slightly decreasing, certain types of complications are increasing.

Patients with BM are becoming increasingly complex patients, and the stress they place on the healthcare system is not limited to surgeons, but will extend to physical therapists, rehabilitation centers, and emergency departments. The increased burden on our healthcare system highlights the need for preparation, optimal patient selection for surgery and promising non-operative strategies. For example, surgical treatment may be prevented by optimizing referral patterns to minimize delay in symptomatic spinal metastases as delayed presentation often lead to surgery.[22,23] More importantly, before choosing invasive and costly surgical strategies, a multidisciplinary oncological team should discuss each patient and formulate an individualized plan while contemplating non-operative treatments. Especially state-of-the-art strategies such as stereotactic radiosurgery need to be carefully considered because they are becoming established options for BM as safe and effective treatments.[24]

Unsurprisingly, patients undergoing surgery for BM had worse clinical outcomes than patients undergoing surgery without BM. This finding likely reflects the poor overall health status and comorbidities of patients with metastatic cancer. Additional factors, such as the immunosuppressive effects of chemotherapy or chronic inflammation associated with cachexia, likely increase adverse events as well.[25] For example, malignancy associated hemostatic changes promote a hypercoagulable state resulting in the increased incidence of embolism as compared with non-BM population.[17,26] In addition, patients in the BM group were more likely to be re-admitted overall, due to embolic and infectious causes, but were less likely to be readmitted for revision surgery. The increased readmission rate for infectious causes may be secondary to the immunosuppressive effects of systemic chemotherapy or due to these patients undergoing longer, more complex surgeries, which are associated with increased infection risk.[27] The decreased readmissions for revision in BM patients was also reported by Park et al., who found a similar decreased readmission rate among patients with metastatic spinal disease compared with patients undergoing spinal surgery for other indications. This decreased readmission rate may be due to decreased survival in the BM group or reluctance to undergo revision surgery due to declining quality of life or health status. Future studies using institutional data should determine an accurate revision rate by correcting for follow-up time and survival.

# CONCLUSION

The number of BM patients undergoing surgery is increasing, inviting the detrimental outcomes of index hospitalization, and further opening the patient to the increased risk of hospital readmission. Choosing the optimal BM candidate for surgical intervention remains difficult as clinicians and patients must weigh the likelihood of improved outcome against the potential for postoperative adverse events for the remaining lifespan. This emphasizes the need to carefully select patients that may benefit from surgical treatment with minimal adverse events. Multidisciplinary approaches are needed to formulate individualized plans while also contemplating non-operative strategies in these patients at high-risk of poor outcomes.

# REFERENCES

1. Manabe J, Kawaguchi N, Matsumoto S, et al. Surgical treatment of bone metastasis: indications and outcomes. *Rev Artic Int J Clin Oncol*. 2005;10:103–111.

2. Coleman RE. Clinical Features of Metastatic Bone Disease and Risk of Skeletal Morbidity. *Clin Cancer Res*. 2006.

3. Quinn RH, Randall RL, Benevenia J, et al. Contemporary management of metastatic bone disease: tips and tools of the trade for general practitioners. *Instr Course Lect*. 2014;63:431–41.

4. Bickels J, Dadia S, Lidar Z. Surgical management of metastatic bone disease. *J Bone Joint Surg Am*. 2009;91(6):1503–1516.

5. Hsiue PP, Kelley BV, Chen CJ, et al. Surgical treatment of metastatic spine disease: an update on national trends and clinical outcomes from 2010 to 2014. *Spine J*. 2020;20(6):915–924.

6. NRD Database Documentation. Healthcare Cost and Utilization Project (HCUP). *Agency Healthc Res Qual Rockville, MD*. 2021.

7. Von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med*. 2007;4(10):1623–1627.

8. Averill RF, Goldfield NI, Muldoon J, et al. A closer look at all-patient refined DRGs. *J AHIMA*. 2002;73(1):46–50.

9. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. 2005;43(11):1130–9.

10. Storesund A, Haugen AS, Hjortås M, et al. Accuracy of surgical complication rate estimation using ICD-10 codes. *Br J Surg*. 2019;106(3):236–244.

11. Consumer Price Index: 2018 in review. *Bur Labor Stat US Dep Labor, Econ. Dly*. 2018.

12. NCSS. *PASS Sample Size Software Cochran-Armitage Test for Trend in Proportions*.

13. Karhade AV, Bongers MER, Groot OQ, et al. Can natural language processing provide accurate, automated reporting of wound infection requiring reoperation after lumbar discectomy? *Spine J*. 2020;20(10):1602–1609.

14. Groot OQ, Bongers MER, Karhade AV, et al. Natural language processing for automated quantification of bone metastases reported in free-text bone scintigraphy reports. *Acta Oncol*. 2020:59(12):1455-1460

15. America I of M (US) C on Q of HC in. Crossing the Quality Chasm. Washington, D.C.: National Academies Press; 2001.

16. Karhade AV, Thio QCBS, Ogink PT, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. *Neurosurgery*. 2019l;85(4):671-681

17. Groot OQ, Ogink PT, Janssen SJ, et al. High risk of venous thromboembolism after surgery for long bone metastases. *Clin Orthop Relat Res*. 2018;476(10):2052–2061.

18. Ghori AK, Leonard DA, Schoenfeld AJ, et al. Modeling 1-year survival after surgery on the metastatic spine. *Spine J*. 2015;15(11):2345–50.

19. Forsberg JA, Eberhardt J, Boland PJ, et al. Estimating survival in patients with operable skeletal metastases: an application of a bayesian belief network. *PLoS One*. 2011;6(5):e19956.

20. Yoshihara H, Yoneoka D. Clinical Study Trends in the surgical treatment for spinal metastasis and the in-

hospital patient outcomes in the United States from 2000 to 2009. *Cancer.* 2014;15;120:901-908

21. Howlader N, Noone AM, Krapcho M, et al. Cancer Statistics Review, 1975-2017 - SEER Statistics. *Nat. Cancer Institute. Bethesda, MD.*

22. van Tol FR, Choi D, Verkooijen HM, et al. Delayed presentation to a spine surgeon is the strongest predictor of poor postoperative outcome in patients surgically treated for symptomatic spinal metastases. *Spine J.* 2019;19(9):1540–1547.

23. van Tol FR, Versteeg AL, Verkooijen HM, et al. Time to surgical treatment for metastatic spinal disease: identification of delay intervals. *Glob spine J.* 2021:2192568221994787.

24. Lutz S, Balboni T, Jones J, et al. Palliative radiation therapy for bone metastases: Update of an ASTRO Evidence-Based Guideline. *Pract. Radiat. Oncol.* 2017;7(1):4–12.

25. Argilés JM, Busquets S, Stemmler B, et al. Cancer cachexia: understanding the molecular basis. *Nat Rev Cancer.* 2014;14(11):754–762.

26. Behranwala KA, Williamson RC. Cancer-associated venous thrombosis in the surgical setting. *Ann Surg.* 2009;249(3):366–375.

27. Ravi B, Jenkinson R, O'heireamhoin S, et al. Surgical duration is associated with an increased risk of periprosthetic infection following total knee arthroplasty: a population-based retrospective cohort study. *EClinicalMedicine.* 2019;16:74–80.

## SUPPLEMENTAL MATERIAL TO CHAPTER 2

**Appendix 1.** ICD10 Procedural and diagnostic codes

**Appendix 2.** Risk factors for outcomes in surgical patients for BM.

**Appendix 3.** Demographics and hospital characteristics of surgical patients with non-BM and BM.

Supplemental material can be consulted online per the website of the journal and/or publisher.

# PART II

# QUALITY OF LIFE AND PHYSICAL FUNCTION

CHAPTER

# 03

# PROSPECTIVE STUDY FOR ESTABLISHING MINIMAL CLINICALLY IMPORTANT DIFFERENCES IN PATIENTS WITH SURGERY FOR LOWER EXTREMITY METASTASES

Michiel E.R. Bongers*, Olivier Q. Groot*, Quirina C.B.S. Thio, Jos A.M. Bramer, Jorrit-Jan Verlaan, Erik T. Newman, Kevin A. Raskin, Santiago A. Lozano-Calderon, Joseph H. Schwab

*Joint first authorship

*Acta Oncologica 2021; Feb 25;1-7*

# ABSTRACT

### Background

The clinical relevance of patient-reported outcomes score changes is often unclear. Especially in patients undergoing surgery due to lower extremity metastases – where surgery is performed in the palliative setting and the goal is to optimize functional mobility, relieve pain and improve overall quality of life.

### Objectives

To assess the minimal clinically important difference (MCID) of Patient-Reported Outcomes Measurement Information System (PROMIS) Pain Interference, Cancer-specific Physical Function, and Global (Physical and Mental Health) in patients treated surgically for impending or completed pathologic fractures.

### Design

Prospective cohort trial with quality-of-life questionnaires.

### Methods

Sixty-five consecutive patients were approached to participate in this study. Six (9.2%) patients declined participation and three (4.6%) did not complete the preoperative surveys due to logistical issues, leaving 56 (86%) patients that completed the preoperative survey. Thirty-three of the 56 (59%) patients completed the postoperative survey: 15 (27%) did not respond or show up at their postoperative consult but were alive at three months, five (9%) died within 1-3 months, and three (5.4%) declined participation. The 33 patients who completed both surveys were included for MCID analysis for whom no missing values were recorded. The MCID of four quality of life questionnaires: PROMIS Pain Interference, PROMIS Cancer-specific Physical Function, and Global (Physical and Mental Health). The anchor-based approach was used to determine the MCID.

### Results

The PROMIS MCIDs (95% confidence interval) for Pain Interference was 7.5 (3.4-12), Physical Function 4.1 (0.6-7.6), Global Physical Health 4.2 (2.0-6.6), and Global Mental Health 0.8 (-4.5-2.9).

### Conclusion

This prospective study successfully defined a MCID for PROMIS Pain Interference of 7.5 (3.4-12), PROMIS Physical Function of 4.1 (0.6-7.6), and Global Physical Health of 4.2 (2.0-6.6) in patients with (impending) pathological fractures due to osseous metastases in the lower extremity; no MCID

could be established for PROMIS Global Mental Health. Defining a narrower MCID value for each subpopulation requires a large, prospective, multicenter study. Nevertheless, the provided MCID values allow guidance to clinicians to evaluate the impact of surgical treatment on a patient's QoL.

# INTRODUCTION

Articulating the potential benefit of surgery can be difficult yet a definition of success after surgery is important to help set expectations of the outcome for the patient, family, and the clinician. This definition of "success" has progressed in the field of orthopaedic surgery from disease centric outcomes such as bony union, joint motion and survival, to patient centric outcomes such as quality of life (QoL) assessments completed by the patient[1]. Quality of life is particularly relevant when the quantity of life is diminished by terminal illness such as patients with (impending) pathological fractures due to lower extremity metastases – where surgery is performed in the palliative and prophylactic setting and the goal is to optimize functional mobility, relieve pain and improve overall QoL[2,3].

The minimally clinical important difference (MCID) is the threshold for when a patient has experienced a minimal clinically relevant change. Thus, the MCID defines that change in an outcome score that is perceived as beneficial[4,5]. For example, a patient changing 1 point on a 10-point VAS score after a treatment may be significant (especially in large trials) but would not be considered clinically important by most patients[6]. The Patient-Reported Outcomes Measurement Information System (PROMIS) Global, PROMIS Cancer-specific Physical Function, and PROMIS Pain Interference are three lower extremity measures that have been scored for a wide variety of oncologic conditions[7,8]. However, no MCIDs are determined for these questionnaires in patients treated surgically for metastatic bone disease of the lower extremity, specifically impending pathologic fractures. As stated earlier, this information is important for this population as surgery is often indicated for palliative purposes. Apart from managing expectations for clinicians and patients during the treatment course, establishing the MCID for PROMIS is expected to aid in the assessment of clinical significance of QoL changes in clinical trials and sample sizes estimates for future studies.

The purpose of the present study was to establish the MCID value of three PROMIS questionnaires - Pain Interference, Cancer-specific Physical Function, and Global (Physical and Mental Health) - in patients treated surgically for impending or completed pathologic fractures secondary to metastatic bone disease of the lower extremity.

# METHODS

### Study Design

This prospective cohort trial was approved by our institutional review board prior to study initiation. All patients attending the orthopaedic oncology clinic in a tertiary hospital were approached for the study in the consecutive months between April 1st 2017 and December 18th 2018. Inclusion criteria for patients were (1) 18 years of age or older, (2) surgical procedure for impending or completed pathologic fracture, and (3) metastatic bone lesion of the lower extremity. Metastatic disease was defined as metastases from solid primary tumors, multiple myeloma, or lymphoma as confirmed by pathology reports[9]; lower extremity was considered as the femur, tibia, and fibula. Exclusion criteria included (1) under 18 years of age, (2) a lack of English proficiency or mental status prohibiting consent for research participation, (3) primary bone lesions, and (4) revision procedures, defined as any subsequent procedure after the index surgery of the pertaining metastatic lesion. Patients to be enrolled in this study would otherwise be undergoing lower extremity surgery, regardless of their participation in this study.

All enrolled subjects were asked to complete the same battery of QoL questionnaires before surgery and one to three months after surgery at their post-operative follow-up appointment. The post-operative survey consisted of the same questionnaires, with the addition of a seven-point Likert global satisfaction anchor question, which was used as the anchor for the assessment of the patients' global perceived effect to calculate the MCID (Appendix 1)[4]. Necessary information on clinical characteristics and QoL was obtained from electronic medical records.

### PROMIS Measures

Three patient reported outcomes were used from PROMIS: (1) PROMIS Pain Interference SF 6a v1.0, (2) PROMIS Global 10 v1.1, and (3) PROMIS Cancer-specific Physical Function v1.1 (Appendix 2.) All these instruments have been previously validated[8]. The PROMIS Pain Interference short form 6a (six items) measures the effects of pain over the last seven days on relevant aspects of one's life completed by the patient. The PROMIS Global Health short form v1.1 (eight items) allows for the assessment of Mental Health (four items) and Physical Health (four items). After previous comparison of several questionnaires for functional outcome in patients with musculoskeletal tumors by our group, the PROMIS Cancer-specific Physical Function was established as the most useful. This was due to "*its reliability over a wide range of ability levels, validity, brevity, and good coverage through computerized adaptive testing (CAT)*"[7]. CAT was used to allow for more efficient physical function assessment[10]. All questionnaire answers were transformed to validated scores with the following ranges: Pain Interference (41-76), Physical Function (15-73), Global Physical Health (16-67), and Global Mental Health (21-67)[8]. A higher score indicates more of the symptoms or health status measured.

### Statistical Analysis

Descriptive statistics were used to characterize the study population. The included patients (those who completed the questionnaires at both timepoints), patients that did not complete the *pre*operative questionnaire, and patients that did not complete the *post*operative questionnaire were compared for baseline differences with the Kruskal-Wallis test for continuous and Fisher's exact test for categorical variables. The following baseline characteristics were compared: gender, age, body mass index, race, Eastern Cooperative Oncology Group (ECOG) performance status established by a clinician at the preoperative and postoperative follow-up visit one to three months after surgery[11], primary tumor, presence of multiple bone metastases, presence of visceral metastases, impending or complete pathologic fracture, history of chemotherapy or local radiotherapy to lower extremity metastasis (yes/no), Katagiri survival score[12], Modified Charlson Comorbidity score[13], and time in years between diagnosis of primary tumor and date of surgery. All variables were collected by a research fellow blinded to the outcome and study population. Graphical examination was used to evaluate performance of the anchor question. Despite being potentially contacted anywhere between one and three months, each patient contributed a single datapoint to the postoperative outcome measure. The questionnaires were administered by REDCap[14]. All statistical analyses were performed using Stata 13.0 (StataCorp LP, College Station, TX, USA).

### MCID methods

The anchor-based approach was used to determine the MCID and 95% confidence intervals (CI) were provided[4,6]. This method compares the mean change between baseline and postoperative outcome measure to a second, external "anchor" question as a reference. On the seven-item anchor question of response to surgery, a response of "a little better" or "better" was considered to be a minimal clinical improvement based on clinical expertise by the treating senior orthopaedic surgeons (Appendix 1). This categorization of the anchor question created two groups of patients experiencing (1) worsening or no change ("much worse", "somewhat worse", "a little worse", and "the same"), and (2) at least minimal improvement ("a little better", "somewhat better", and "much better") based on the anchor question after surgery. Naturally, only questionnaires with complete pre- and postoperative data can be included in this method. To provide supporting data that the MCID determined by the anchor-based approach exceeded the measurement error, the anchor-based results were compared to the mean baseline and mean postoperative standard error of measurement (SEM)[15,16]. The SEM is a statistical measure that represents the reliability of patients' scores on a QoL assessment tool – a patient rating the depression level as 7 of 10 and later as 8 of 10, without any actual change in the perceived depression level. The calculated anchor-based MCID value being higher than this SEM, reflects that the MCID value is not due to chance or random variation, but due to a real change (e.g. surgical treatment).

To account for the sample and method dependent variations, we provided a range of MCID values with corresponding percentages of the total score in the supplementary material by including the distribution-based approach[17], rather than an absolute single threshold from the anchor-based method[18]. Both the anchor and the distribution-based approach have limitations and no consensus exists on the preferred methodology[6]. Therefore, both results are provided in the supplemental material until better uniform guidelines exists for identifying an appropriate MCID analysis. Sample size limited a sub analysis for whether MCID estimated values differed according to baseline clinical or demographic characteristics.

# RESULTS

Sixty-five consecutive patients were approached to participate in this study. Six (9.2%) patients declined participation and three (4.6%) did not complete the *pre*operative surveys due to logistical issues, leaving 56 (86%) patients that completed the *pre*operative survey. Thirty-three of the 56 (59%) patients completed the *post*operative survey: 15 (27%) did not respond or show up at their postoperative consult but were alive at three months, five (9%) died within 1-3 months, and three (5.4%) declined participation. The 33 patients who completed both surveys were included for MCID analysis for whom no missing values were recorded (Figure 1).

The median patient age was 68 years (interquartile range [IQR] 59 – 70, Table 1). The indication for surgery was an impending pathologic fracture in 27 (82%) and pathologic fracture in 6 (18%). Additional analyses for baseline characteristics were provided between the non-*pre*operative responders (n=9), non-*post*operative responders (n=23), and included patients (n=33). Non-responders had a higher Katagiri survival score and lower three- and six-month survival rates (Appendix 3).

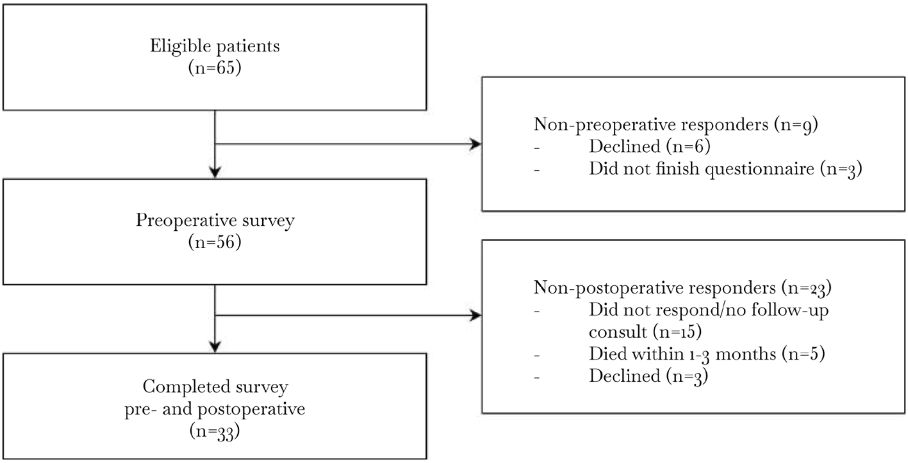

Eligible patients
(n=65)
Non-preoperative responders (n=9)
- Declined (n=6)
- Did not finish questionnaire (n=3)
Preoperative survey
(n=56)
Non-postoperative responders (n=23)
- Did not respond/no follow-up consult (n=15)
- Died within 1-3 months (n=5)
- Declined (n=3)
Completed survey
pre- and postoperative
(n=33)

**Figure 1.** Flowchart of enrolled patients.

**Table 1:** Sociodemographic and clinical characteristics of the included patients (n 33)

| *Variables* | *Median (IQR)* |
|---|---|
| Age | 68 (59-70) |
| BMI | 27 (24-31) |
| Modified Charlson Comorbidity Index | 6 (6-6) |
| Duration of primary diagnosis until metastatic operation (years) | 2.8 (0.5-8.2) |
|  | n (%) |
| Men | 14 (42) |
| Ethnicity |  |
| Caucasian | 32 (97) |
| Hispanic | 1 (3.0) |
| Preoperative ECOG score |  |
| 0-2 | 29 (88) |
| 3-4 | 4 (12) |
| Primary tumor |  |
| Breast cancer | 7 (21) |
| Kidney cancer | 6 (18) |
| Lung cancer | 4 (12) |
| Thyroid cancer | 2 (6.1) |
| Melanoma | 2 (6.1) |
| Myeloma | 2 (6.1) |
| Other[a] | 10 (30) |
| Visceral metastases |  |
| Yes (lung, liver, or brain) | 15 (45) |
| No | 18 (55 |
| Multiple bone metastases |  |
| Yes | 17 (52) |
| No | 16 (48) |
| Previous local radiotherapy |  |
| Yes | 5 (15) |
| No | 28 (85) |
| Previous systemic therapy |  |
| Yes | 23 (70) |
| No | 10 (30) |
| Katagiri survival score[b] |  |
| 0-3 | 18 (55) |
| 4-7 | 14 (42) |
| 6-10 | 1 (3.0) |
| Fracture type |  |
| Impending | 27 (82) |
| Pathologic | 6 (18) |
| Fracture location |  |
| Femur | 31 (94) |
| Tibia | 2 (6.1) |
| Operative treatment type |  |
| Intramedullary nailing | 17 (52) |
| Endoprosthetic reconstruction | 13 (39) |
| Dynamic hip screw | 2 (6.1) |
| Plate-screw fixation | 1 (3.0) |

*Continued on next page*

| *Postoperative* | |
|---|---|
| Reoperation | 2 (6.1) |
| ECOG at 1-3 months | |
| Worsened | 3 (9.1) |
| Same | 20 (61) |
| Improved | 10 (30) |
| Mobility at 1-3 months | |
| Unassisted | 8 (24) |
| Cane/crutch | 3 (9.1) |
| Walker | 22 (67) |
| Survival[c] | |
| 3-months | 31 (94) |
| 6-months | 24 (73) |

*IQR=interquartile range; BMI=body mass index; ECOG=Eastern Cooperative Oncology Group*

*a This category includes colorectal cancer 1 (3.0%), cholangiocarcinoma 1 (3.0%), chordoma 1 (3.0%), lymphoma 1 (3.0%), epithelioma 1 (3.0%), ovarian cancer 1 (3.0%), prostate cancer 1 (3.0%), head and neck cancer 1 (3.0%), sarcoma 1 (3.0%), and unknown 1 (3.0%).*

*b Survival prediction at 12 months for patients with bone metastases: low-risk group (score of ≤3), survival rate >80%; the intermediate-risk group (score of 4–6), survival rate 30–80%; and the high-risk group (score of 7–10), survival rate ≤10%.*

*c Loss to follow-up at 3 months was 0 patients, and at 6 months 2 patients (6.1%)*

### MCIDs

For the 33 patients that completed both *pre-* and *post*operative survey, 70% patients (n=23) reported “much better”, 15% patients (n=5) “somewhat better”, 9.1% patients (n=3) “a little better”, 3.0% (n=1) patients “a little worse”, and 3.0% patients (n=1) reported being “much worse” at 1-3 months after surgery based on the anchor question. This corresponds with 31 patients (94%) reporting at least minimal improvement (e.g., “a little better” or better) and two patients (6.1%) reporting deterioration. The score distribution for the questionnaires at both time points, the mean change between the pre- and postoperative questionnaire, and the anchor-based MCID were determined (Table 2).

The MCID values calculated with the anchor-based approach in patients undergoing surgery for an (impending) pathological fracture due to lower extremity bone metastases were: PROMIS Pain Interference 7.5 [95% CI: 3.4 – 12], PROMIS Physical Function 4.1 [95% CI: 0.6 – 7.6], PROMIS Global Physical Health 4.3 [95% CI: 2.0 – 6.6], and PROMIS Global Mental Health 0.8 [95% CI: -4.5 – 2.9]. All MCIDs values, except Global Mental Health, were greater than the mean of SEM both at preoperative and postoperative assessment, indicating that the estimate of MCID exceeded measurement error. In other words, the established MCIDs, except Global Mental Health, in the PROMIS scores were deemed to be due to surgical treatment, not random variation. Combining the anchor-based and distribution-based approach, MCID ranges were as follows: PROMIS-Pain Interference, 2.9 to 7.5 points (8.3-21% of the total score); PROMIS-Physical Function, 2.5 to 4.2 points (4.3-7.2% of the total score); PROMIS Global Physical Health, 2.1 to 5.9 points (4.0-11% of the total score); and PROMIS Global Mental Health, 0.8 to 6.0 points (1.7-13% of the total score; Appendix 4).

**Table 2.** Comparison of pre- and postoperative scores and MCID estimates.

| *QoL questionnaires* | *Mean (SD)* | | *Mean change* | *Anchor-based MCID* |
|---|---|---|---|---|
| | *Preoperative* | *Postoperative* | *(SD)* | *(95% CI)* |
| PROMIS Pain Interference [a] | 65 (8.6) | 58 (9.1) | -7.4 (11) | 7.5 (3.4-12) |
| PROMIS Physical Function [b] | 30 (7.4) | 34 (8.5) | 4.5 (9.4) | 4.1 (0.6-7.6) |
| PROMIS Global Physical [b] | 35 (6.9) | 39 (7.7) | 4.4 (6.2) | 4.3 (2.0-6.6) |
| PROMIS Global Mental Health [b] | 47 (7.6) | 46 (9.1) | -1.0 (9.7) | 0.8 (-4.5-2.9) |

*MCID=minimal clinically important difference; QoL=quality of life; SD=standard deviation; CI=confidence interval; PROMIS=Patient-Reported Outcomes Measurement Information System. No missing values were recorded. The pre- and postoperative mean are based on 33 patients; and the anchor based MCID on 31 patients (the patients that improved based on the anchor question)*
*a: Higher score represents a worse health status; a greater degree of the pain interference symptoms is present.*
*b: Higher score represents better health status.*

# DISCUSSION

Patients undergoing surgery for impending or pathologic fracture of lower extremity bone metastases often have poor QoL due to invalidating ambulatory status and pain[2]. Understanding the MCID allows the determination of the benefits of surgical treatments and interpretation of group-level data as opposed to assessing individual patients when conducting research. The anchor-based approach determined that a decrease of the following values corresponds to a MCID: PROMIS-Pain Interference 7.5 [95% CI: 3.4 – 12], PROMIS-Physical Function 4.1 [95% CI: 0.6 – 7.6], and PROMIS Global Physical Health 4.3 [95% CI: 2.0 – 6.6]. No MCID could be established for PROMIS Global Mental Health. To our knowledge, this is the first study establishing the MCID for three PROMIS questionnaires in patients surgically treated for (impending) pathologic fractures due to osseous metastases of the lower extremity[6].

This study has certain inherent limitations. First, the sample size is relatively small, though most prospective skeletal metastases series are small because of the low incidence and poor survival[19,20]. This is noted in the wide confidence interval, which is comparable with similar sample size studies including oncologic patients[21,22]. Analysis of a larger, multicenter cohort can narrow this interval down. Despite the wide interval of MCIDs, a MCID greater than the SEM was observed in three of the four questionnaires. This reflects that the MCID is not due to chance or random variation, but due to a real change (e.g., surgical treatment). One may argue that this real change can also be due to confounding factors and not surgical treatment. The sample size limited a sub analysis for whether MCID estimated values differed according to baseline clinical or demographic characteristics. For example, completed pathological fractures are known to be associated with worse outcomes as compared to impending fractures[23,24]. This may correspond with smaller MCIDs as these patients experience less postoperative improvement in pain and mobility as compared to impending pathologic fractures.

Second, 32 of the 65 (49%) eligible patients declined, had incomplete data, were lost to follow-up, or died, and this could introduce bias into our results (Appendix 3). A retrospective analysis of the causes of patients not completing the surveys preoperative or postoperative demonstrated that their Katagiri survival score and three and six-month survival rate were worse than the included patients; the other characteristics such as preoperative ECOG score, ECOG improvement, and mobility at one to three months were comparable. These non-respondents may have experienced more severe postoperative deterioration compared with the included patients, and this could explain our low percentage of non-improved patients (6.1%; 2 of 33). However, this did not affect our anchor-based MCID analysis since this approach only takes into account the patients that report an improvement on the anchor question[4]. Despite having no effect on our MCID analysis, the low number of non-improved patients withheld us from using a receiver-operating characteristics curve to develop a MCID threshold. However, differences were found between three of the four MCID values and the mean of SEM, which supports validity of our MCID values since the MCIDs estimates exceeded measurement error[4].

Third, the choice of follow-up intervals may have affected the MCID value - was to three months an appropriate timepoint to determine MCID for metastasis surgery, as compared to say, for example six or 12 months? Since patients with extremity metastatic disease undergoing surgery have a three-month survival rate of 50%, we chose the follow-up intervals at one to three months to minimize loss to follow-up prior to potential marked improvement and allow for a rehabilitation period following the surgery[19,20]. Also, we believe one to three months is an appropriate time point as these patients need to see improvement in a short time horizon that is proportional to their life expectancies and a recent study reports that QoL is restored within 6 weeks in patients with femoral metastases undergoing endoprosthesis[25]. Additional analysis showed no difference between the patients who completed the postoperative scores at one month and three months in different amounts of improvement on PROMIS scores. Fourth, this was a tertiary institution study, in a single region of the country, and may not reflect practice in other regions. No uniform guidelines exist for treating impending or completed pathological fractures and this can result in different surgical techniques, timing of treatment, and philosophy on whether or not to treat in this complicated patient population. Fifth, an anchor-based approach was used to determine the MCID, which is open to criticism[26]. The rating scale of the anchor question requires an arbitrary choice of cut-off points to define degrees of clinical improvement and may be affected by recall bias[27]. It also does not take into account the statistical characteristics of a group's baseline, which the distribution-based approach does[4]. At present, both the anchor and the distribution-based approach have limitations and no consensus exists on the preferred methodology[6]. Both results are provided in the supplemental material until better uniform guidelines exists for identifying an appropriate MCID analysis. However, we recommend using the anchor-based results to ensure consistent application of the MCID in practice and homogeneity between studies[4,15,28,29].

Only one other similar study by Yost et al.[15] established MCIDs in the 10-item PROMIS Physical and Pain Interference for patients with advanced cancer-stage (III and IV) undergoing chemotherapy, radiation therapy, or both. This study included 101 patients with 23 clinically relevant, self-reported anchors and determined that a 4.0-6.0 change in both questionnaires represented clinically meaningful change. Although most similar to this study, the range of change in Yost et al.[15] is lower than our current findings. This could be explained due to differences in clinical anchors used, number of included patients, different item versions of the questionnaires and patient population (i.e. severity of treatment; surgical treatment versus chemotherapy and/or radiation). The MCIDs in this study for the pain and physical surveys were relatively high, which is reflected by the fact that these patients are experiencing a major decrease in mobility and an increase in pain interference scores at baseline due to their (impending) pathologic fracture[2]. Thus, by performing surgical stabilization the impact on physical function and pain interference can greatly improve, giving the large changes that were observed in this cohort. This would be beneficial in a patient population that has a 3-month survival rate of 50% as immobility and pain are the greatest concerns on patients' minds[1,20,30]. Also, the MCID values may be narrowed towards a smaller value in the wide confidence interval range when the sample size is increased.

The anchor-based approach did not yield a usable estimate for the MCID of the PROMIS Global Mental Health. The MCID estimates included zero and was not different than the SEM. This may be explained because changes in mental QoL often occur in the earlier phase of diagnosis[31]. Over time, patients can adapt to their disease, which can contribute to the stabilization of mental QoL scores over time regardless of a treatment or not[15]. In particular, patients undergoing surgical treatment for bone metastases are in a more advanced stage of their disease and their mental health status could have changed – to their current mental status – in an earlier stage of the disease course (patients had a median of 2.5 years (IQR 0.5-8.2) from primary tumor diagnosis until the surgery related to this study)[31]. Furthermore, the included patients often had concurrent treatments for various other sequelae of their primary or metastatic disease, which could have impacted their mental status when contributing to the surveys. As a result, the favorable outcome of the surgery in terms of physical function and pain interference related to this study may not be reflected in a changed mental status, which could explain the non-detectable MCID. This has been observed in prior oncology studies where despite less pain and overall better QoL, patients experience similar mental health *pre-* and *post*treatment[32,33].

Future research – to address the above limitations - should include a larger sample from a multicenter cohort and provide stratified QoL results by demographic and clinical characteristics such as primary tumor type, preoperative ECOG and comorbidities to elucidate a more definite MCID for each subpopulation. Nevertheless, this prospective study established valid and usable MCIDs through a rigorous anchor-based method. To our knowledge, this is the first prospective study determining MCIDs in patients treated surgically for impending or completed pathologic fractures secondary to metastatic bone disease of the lower extremity. As recently recommended by Janssen[34], prospective studies such as this current study that "include homogenous samples like lower-extremity bone metastases, and assess multiple QOL and physical function questionnaires to calculate corresponding MCIDs, focusing on anchor-based MCIDs" are necessary to improve and evaluate treatment goals in this complicated patient population. By using the provided MCIDs as benchmark, this study provides valuable information in managing expectations for clinicians and patients during the treatment course and the assessment of clinical significance of QoL changes in clinical trials.

# CONCLUSION

This prospective study successfully defined a MCID for PROMIS Pain Interference of 7.5 (3.4-12) points, PROMIS Physical Function of 4.1 (0.6-7.6) points, and Global Physical Health of 4.2 (2.0-6.6) points in patients with impending or complete pathological fractures due to osseous metastases in the lower extremity; no MCID could be established for PROMIS Global Mental Health. These MCID values can be used for managing expectations for clinicians and patients during the treatment course, interpreting group-level data as opposed to assessing individual patients when conducting research and aid in the assessment of clinical significance of QoL changes in clinical trials. Future research should include a large, multicenter and provide stratified MCID results by demographic and clinical characteristics such as primary tumor type, fracture type and comorbidities in order to elucidate a more definite MCID for each subpopulation.

# REFERENCES

1. Cleeland CS, O'Mara A, Zagari M, et al. Integrating pain metrics into oncology clinical trials. *Clin Cancer Res.* 2011;17(21):6646–6650.

2. Schwab JH. CORR Insights®: Early improvement in pain and functional outcome but not quality of life after surgery for metastatic long-bone disease. *Clin Orthop Relat Res.* 2018;476(3):546–547.

3. Janssen SJ, Teunis T, Hornicek FJ, et al. Outcome after fixation of metastatic proximal femoral fractures: A systematic review of 40 studies. *J Surg Oncol.* 2016;114(4):507–519.

4. Sedaghat AR. Understanding the minimal clinically important difference (MCID) of patient-reported outcome measures. *Otolaryngol Head Neck Surg. (United States).* 2019;161(4):551–560.

5. Bedard G, Zeng L, Lam H, et al. Meaningful change in oncology quality-of-life instruments: A systematic literature review. *Expert Rev Pharmacoeconomics Outcomes Res.* 2012;12(4):475–483.

6. Copay AG, Eyberg B, Chung AS, et al. Minimum clinically important difference: current trends in the orthopaedic literature, part II: lower extremity: a systematic review. *JBJS Rev.* 2018;6(9):e2.

7. Janssen SJ, Paulino Pereira NR, Raskin KA, et al. A comparison of questionnaires for assessing physical function in patients with lower extremity bone metastases. *J Surg Oncol.* 2016;114(6):691–696.

8. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH roadmap cooperative group during its first two years. *Med Care.* 2007;45(5).

9. Nathan SS, Healey JH, Mellano D, et al. Survival in patients operated on for pathologic fracture: implications for end-of-life orthopedic care. *J Clin Oncol.* 2005;23(25):6072–6082.

10. Rose M, Bjorner JB, Gandek B, et al. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *J Clin Epidemiol.* 2014;67(5):516–526.

11. Oken M, Creech R, Tormey D, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol.* 1982;5(6):649–656.

12. Katagiri H, Okada R, Takagi T, et al. New prognostic factors and scoring system for patients with skeletal metastasis. *Cancer Med.* 2014;3(5):1359–1367.

13. Quan H, Li B, Couris CM, et al. Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol.* 2011;173(6):676–682.

14. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)- a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009;42(2):377–81.

15. Yost KJ, Eton DT, Garcia SF, et al. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *J Clin Epidemiol.* 2011;64(5):507–516.

16. de Vet HCW, Terluin B, Knol DL, et al. Three ways to quantify uncertainty in individually applied "minimally important change" values. *J Clin Epidemiol.* 2010;63(1):37–45.

17. Beckerman H, Roebroeck ME, Lankhorst GJ, et al. Smallest real difference, a link between reproducibility and responsiveness. *Qual Life Res.* 2001;10(7):571–578.

18. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life

research: How meaningful is it? *Pharmacoeconomics*. 2000;18(5):419–423.

19. Karhade AV, Thio QCBS, Ogink PT, et al. **Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation.** *Neurosurgery*. 2019:1;85:671-681

20. Thio QCBS, Karhade AV, Ogink PT, et al. **Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease.** *Clin Orthop Relat Res.* 2020;478(2):322–333.

21. Zuckerman SL, Chotai S, Devin CJ, et al. **Surgical resection of intradural extramedullary spinal tumors: Patient reported outcomes and minimum clinically important difference.** *Spine (Phila. Pa. 1976).* 2016;41(24):1925–1932.

22. Wong E, Zhang L, Kerba M, et al. **Minimal clinically important differences in the EORTC QLQ-BN20 in patients with brain metastases.** *Support Care Cancer*. 2015;23(9):2731–2737.

23. Franis KC. **Prophylactic internal fixation of metastatic osseous sesions.** *Cancer*. 1960;13:75-76

24. Harrington KD. **New trends in the management of lower extremity metastases.** *Clin Orthop Relat Res.* 1982;(169):53–61.

25. Sørensen MS, Horstmann PF, Hindsø K, et al. **Use of endoprostheses for proximal femur metastases results in a rapid rehabilitation and low risk of implant failure. A prospective population-based study.** *J Bone Oncol.* 2019;19:100264.

26. Hägg O, Fritzell P, Nordwall A. **The clinical importance of changes in outcome scores after treatment for chronic low back pain.** *Eur Spine J.* 2003;12(1):12–20.

27. Van Der Roer N, Ostelo RWJG, Bekkering GE, et al. **Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain.** *Spine (Phila. Pa. 1976).* 2006;31(5):578–582.

28. Sorensen AA, Howard D, Tan WH, et al. **Minimal clinically important differences of 3 patient-rated outcomes instruments.** *J Hand Surg Am.* 2013;38(4):641–649.

29. Terwee CB, Roorda LD, Dekker J, et al. **Mind the MIC: large variation among populations and methods.** *J Clin Epidemiol.* 2010;63(5):524–534.

30. Engelmann D, Scheffold K, Friedrich M, et al. **Death-related anxiety in patients with advanced cancer: validation of the german version of the death and dying distress scale.** *J Pain Symptom Manage.* 2016;52(4):582–587.

31. Yost KJ, Sorensen MV, Hahn EA, et al. **Using multiple anchor- and distribution-based estimates to evaluate clinically meaningful change on the functional assessment of cancer therapy-biologic response modifiers (FACT-BRM) instrument.** *Value Heal.* 2005;8(2):117–127.

32. Dea N, Versteeg AL, Sahgal A, et al. **Metastatic spine disease: should patients with short life expectancy be denied surgical care? An international retrospective cohort study.** *Neurosurgery*. 2020:1;87:303-311

33. Poghosyan H, Sheldon LK, Leveille SG, et al. **Health-related quality of life after surgical treatment in patients with non-small cell lung cancer: A systematic review.** *Lung Cancer*. 2013;81(1):11–26.

34. Janssen SJ. CORR Insights®: **What are the minimum clinically important differences in SF-36 scores in patients with orthopaedic oncologic conditions?** *Clin Orthop Relat Res.* 2020;478(9);2159-2160

# SUPPLEMENTAL MATERIAL TO CHAPTER 3

**Appendix 1.** Anchor question.

**Appendix 2.** Baseline comparison between included non-postoperative, and non-preoperative group.

**Appendix 3.** Quality of life questionnaires that were used in this prospective study.

**Appendix 4.** Distribution-based and anchor-based MCID estimates.

Supplemental material can be consulted online per the website of the journal and/or publisher.

CHAPTER

04

# QUALITY OF LIFE CHANGES AFTER OPEN SURGERY FOR METASTATIC SPINAL DISEASE: A SYSTEMATIC REVIEW AND META-ANALYSIS

Nuno R. Paulino Pereira, Olivier Q. Groot, Jorrit-Jan Verlaan, Michiel E.R. Bongers, Peter K. Twining, Neal D. Kapoor, Cornelis N. van Dijk, Joseph H. Schwab, Jos A.M. Bramer

*Clinical Spine Surgery 2021*

# ABSTRACT

### Background

It remains questionable to what extent open surgery improves quality of life (QoL) for metastatic spinal disease, it would be interesting to quantify the magnitude and duration of QoL benefits – if any – after surgery for spinal metastases.

### Objectives

To assess QoL after open surgery for spinal metastases, and how surgery affects physical, social/family, emotional, and functional well-being.

### Design

Systematic review and meta-analysis.

### Methods

A literature search was performed in PubMed, Embase, and the Cochrane library from inception to February 6th 2020, and used synonyms for 'spine', 'metastatic', and all questionnaires that have been suggested to measure QoL in patients with spinal metastases. Included were studies measuring QoL before and after non-percutaneous, open surgery for spinal metastases for various indications including pain, spinal cord compression, instability or tumor control. A random-effect model assessed standardized mean differences (SMD) of summary QoL scores between baseline and 1, 3, 6, or 9-12 months after surgery.

### Results

The review yielded 10 studies for data extraction. The pooled QoL summary score improved from baseline to 1-month (SMD=1.09, p<0.001), to 3-months (SMD=1.28, p<0.001), to 6-months (SMD=1.21, p<0.001), and to 9-12 months (SMD=1.08, p=0.001). Surgery improved physical well-being during the first 3-months (SMD=0.94, p=0.022), improved emotional (SMD=1.19, p=0.004) and functional well-being (SMD=1.08, p=0.005) during the first 6-months, and only improved social/family well-being at month 6 (SMD=0.28, p=0.001).

### Conclusion

Patients with spinal metastases undergoing surgery experienced improved QoL, and rapidly improved physical, emotional, and functional well-being; it had minimal effect on social/family well-being. However, choosing the optimal candidate for surgical intervention in the setting of spinal metastases remains paramount: otherwise, postoperative morbidity and complications may

outbalance the intended benefits of surgery. Future research should report clear definitions of selection criteria and surgical indication and provide stratified QoL results by indication and clinical characteristics such as primary tumor type, preoperative Karnofsky and Bilsky scores to elucidate the optimal candidate for surgical intervention.

# INTRODUCTION

The spine is the most common location for bone metastases.[1–3] Surgery can be offered to carefully selected patients, with the goal to improve survival, relieve pain symptoms, restore spinal stability, counteract neurologic deficits, and improve, or at least maintain, quality of life (QoL).[4–6]

QoL is a term that refers to "a person's sense of well-being that stems from satisfaction or dissatisfaction with the areas of life that are important to him or her."[7] Surgeons are now commonly measuring QoL in patients who undergo surgery for metastatic spinal disease to determine how much surgery improves QoL in this advanced life phase.[8–10] Adequately measuring QoL in this population is challenging due to a high loss to follow-up after surgery, the influence of comorbidities, and confounding psychosocial and emotional issues that might already exist before their disease.[11–13] Furthermore, studies evaluating QoL in patients with metastatic spinal disease have used many different questionnaires, which may not be validated in this patient population, making it difficult to compare study results. It would be interesting to quantify the magnitude and duration of QoL benefits after surgery for metastatic spinal disease.

A systematic review and meta-analysis were conducted to assess QoL after surgery for metastatic spinal disease. Secondarily, this study assessed how surgery affects physical, social/family, emotional, and functional well-being by quantifying subdomains of QoL questionnaires.

# METHODS

### Literature Search and Study Selection

We report our results per the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, with a PRISMA checklist and algorithm.[14]A literature search was performed in PubMed, EMBASE, and the Cochrane library from inception to February 6th 2020, and used synonyms for 'spine', 'metastatic', and all questionnaires that have been suggested to measure QoL in patients with metastatic spinal disease (Appendix 1).[12,13]

Studies were included that measured QoL before and after surgery for metastatic spinal disease for various indications including pain, spinal cord compression, instability, and tumor control. We included any type of non-percutaneous, open surgical treatment (e.g., decompression, corpectomy, stabilization, reconstruction, or debulking, or any combination thereof). Exclusion criteria were: (1) studies where vertebroplasty, kyphoplasty, or radiosurgery were the only procedures performed (2) studies without separate data for patients surgically treated for metastatic spinal disease, (3) non-English studies, (4) affiliated studies with high republication suspicion, and (5) non-relevant study types –such as reviews, case-reports, animal studies, technique papers, letters to editors, and meeting abstracts.

Two researchers (NRPP, OQG) independently screened titles and abstracts from 992 studies that were identified through the literature search: 72 studies were potentially eligible and full-text screened by each of the two researchers (Figure 1). After reading full-text studies, 56 were excluded because these were abstracts from meetings (n=26), did not mention a separate analysis for the target patient population (n=18), were non-English (n=6), had vertebroplasty or kyphoplasty as only surgical procedure (n=3), or high republication suspicion (n=3, included was the most recent of the two studies). The two researchers each read bibliographies of the 16 included studies and identified one study that had been missed by the initial search. The same two researchers discussed all uncertainties and inconsistencies until consensus was obtained for study selection, data extraction, and critical appraisal.

### Critical Appraisal

Each of two reviewers (NRPP, OQG) independently appraised the methodological quality of the 17 included studies by identical predefined extraction sheets, based on the methodological index for non-randomized studies (MINORS) criteria.[15] These MINORS criteria were created and validated by a group of surgeons to assess the methodological quality of non-randomized observational studies; they contain eight items for observational non-comparative studies, and each item is scored 0 (not reported), 1 (reported but inadequate) or 2 (reported and adequate). Items 5, 6, 7, and 8 were rephrased to make them applicable for our systematic review; the patient completed the quality of life instrument (item 5), excluded studies did not obtain questionnaires after at least 6 months (item 6), studies where less than 60% of patients were measured at 6 months' follow-up (item 7), and mentioning of sample size calculation (item 8, Table 1). Studies that did not reach a full score (i.e. 2 points) on item 6 or item 7, or studies with less than 10 points on the total score were excluded for further data extraction.

### Data Extraction

The following baseline variables were extracted from included studies by predetermined data extraction sheets: author, year, countries where patients were included, study type, comparison group (if applicable), number of surgical patients, age and gender distribution, primary cancer types, regions in the spine operated on, surgical treatments, number of patients that completed the questionnaire(s) per time point, and complications reported. We extracted the following outcome variables from the included studies: type of QoL questionnaire(s) included, QoL summary score per time point, subdomain scores of QoL questionnaire(s) per time point (if applicable), and conclusion regarding the QoL measurement.

We contacted authors from three studies to obtain exact means and standard deviations (SD) for QoL scores;[8,9,16] two provided us with the requested data[9,16] and one did not –therefore we could not include the latter for meta-analytic purposes.[8]

### Included Questionnaires

Six different QoL questionnaires were mentioned in the studies that were included for data extraction (Appendix 2). We only included the EuroQol 5 Dimensions (EQ5D) and Functional Assessment of Cancer Therapy – General (FACT-G) for meta-analytic purposes. Data for other questionnaires were incomplete or did not provide a QoL summary score (e.g. Short Form 36).

The EQ5D is a generic questionnaire –i.e. not disease specific—and assesses QoL through 5 items where higher scores indicate better QoL.[17] The FACT-G questionnaire was specifically designed to assess QoL in cancer patients, and contains 27 items. The total score can range between 0 and 108, and higher scores indicate better QoL.[18] In addition, the FACT-G measures the following 4 subdomains: physical well-being (7 items), social/family well-being (7 items), emotional well-being (6 items), and functional well-being (7 items).

### Statistical Analysis

We used a random-effect model to calculate standardized mean differences (SMD) between summary QoL scores from baseline to either 1, 3, 6, or 9-12 months –we fused the 9 and 12-month time points into one time point to retain statistical power as most studies reported either one of the two endpoints. We pooled the EQ5D and FACT-G summary scores together as these both provide QoL summary scores (pooling summary scores of different questionnaires is common[19,20]), and rescaled summary scores to make them equal.[21,22] A SMD of +1 signifies that the QoL summary score is 1 SD better on that time point, when compared with the baseline score. We also used a random-effect model to calculate SMD's for physical, social/family, emotional, and functional well-being FACT-G subdomains from baseline to either 1, 3, 6, or 9-12 months. We visualized changes in QoL or subdomain scores over time with longitudinal plots, depicted the included studies per graph and included a line for the pooled SMD's.

We assessed consistency of results with the $I^2$ and $Tau^2$ statistic for QoL summary scores. $I^2$ gives the percentage of variation across studies due to heterogeneity, and ranges from 0-100% –higher percentages indicate more heterogeneity across studies.[23] $Tau^2$ provides between study variance in a

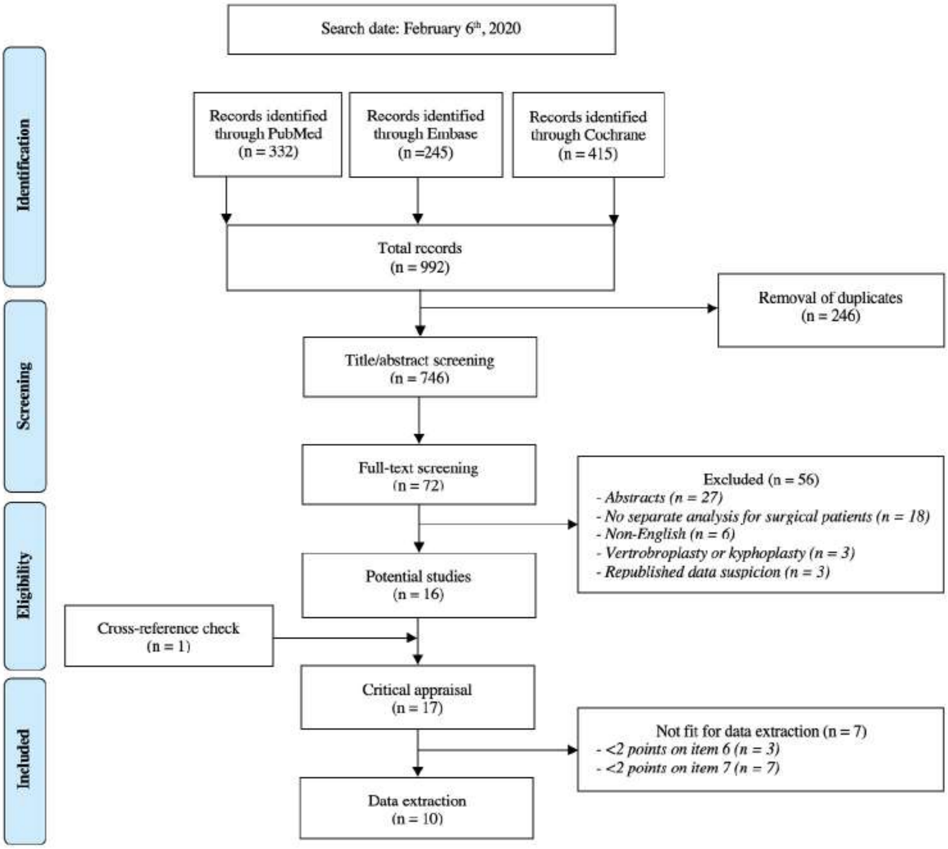

Search date: February 6th, 2020
Identification
Screening
Eligibility
Included
Records identified through PubMed (n = 332)
Records identified through Embase (n =245)
Records identified through Cochrane (n = 415)
Total records (n = 992)
Removal of duplicates (n = 246)
Title/abstract screening (n = 746)
Full-text screening (n = 72)
Excluded (n = 56)
- Abstracts (n = 27)
- No separate analysis for surgical patients (n = 18)
- Non-English (n = 6)
- Vertrobroplasty or kyphoplasty (n = 3)
- Republished data suspicion (n = 3)
Potential studies (n = 16)
Cross-reference check (n = 1)
Critical appraisal (n = 17)
Not fit for data extraction (n = 7)
- <2 points on item 6 (n = 3)
- <2 points on item 7 (n = 7)
Data extraction (n = 10)

**Figure 1.** Flow chart of the literature search and study selection.

random-effects meta-analysis, and scores greater than 1 indicate substantial statistical heterogeneity.[24] We used Stata® 14.0 (StataCorp LP, College Station, TX, USA) for statistical analyses, Mendeley Desktop Version 1.19.4 (Mendeley Ltd., London, UK) as a reference management software, and considered two-tailed P-values less than 0.05 to be significant. No funding was received.

# RESULTS

## Study Selection and Study Characteristics

Seven studies were excluded because they did not obtain questionnaires after at least 6 months (n=3) and/or had less than 60% follow-up at 6 months (n=7) (Table 1). Ten prospective studies were selected for data extraction.[8–10,16,21,22,25–28] The number of surgical patients in these studies varied from 29 to 922, mean ages ranged from 49 to 69 years, and patients were more often of the male gender

Table 1. Modified version of the methodological index for non-randomized studies (MINORS criteria) and scoring for the studies selected for critical appraisal.

| *Methodological items for non-randomized studies*[a] | *Interpretation* | *Score*[b] |
|---|---|---|
| 1. A clearly stated aim | The question addressed is precise and relevant in the light of available literature. | 0 - 2 |
| 2. Inclusion of consecutive patients | All patients satisfying the criteria for inclusion have been included in the study during the study period. | 0 - 2 |
| 3. Prospective collection of data | Data were collected according to a protocol established before the beginning of the study. | 0 - 2 |
| 4. Endpoints appropriate to the aim of the study[c] | The authors clearly elaborate how they scored the quality of life instrument, and what version and language they use. | 0 - 2 |
| 5. Unbiased assessment of the study endpoint[c] | The patient completed the quality of life instrument. | 0 - 2 |
| 6. Follow-up period appropriate to the aim of the study[c] | Follow-up for should be at least 6 months to allow assessment of main endpoint and possible adverse events. | 0 - 2 |
| 7. Loss to follow up[c] | ≥ 60% of patients measured at baseline should be measured at 6 months follow-up (excluding deceased patients in denominator) | 0 - 2 |
| 8. Prospective calculation of the study size[c] | Detailed information on an adequate sample size calculation or post-hoc power calculation. | 0 - 2 |

| *Methodological items for non-randomized studies*[a] | Studies that were selected for critical appraisal (n=17) | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Morgen | Choi | Wu | Tangpat | Zhang | Quan | Tang | Fehlings | Zheng | Falicov | Miscuzi | Mannion | Konovalov | De Ruiter | Versteeg | Miyazaki | Ma |
| 1. A clearly stated aim | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 |
| 2. Inclusion of consecutive patients | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 2 | 2 |
| 3. Prospective collection of data | 2 | 2 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 0 | 1 | 0 | 2 | 2 | 2 | 2 |
| 4. Endpoints appropriate to the aim of the study | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 |
| 5. Unbiased assessment of the study endpoint[c] | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 2 |
| 6. Follow-up period appropriate to the aim of the study[c] | 2 | 2 | 2 | 2 | 1[d] | 2 | 2 | 2 | 2 | 2 | 1[d] | 1[d] | 2 | 2 | 2 | 2 | 2 |
| 7. Loss to follow up[c] | 2 | 2 | 2 | 1[d] | 1[d] | 1[d] | 2 | 2 | 2 | 2 | 1[d] | 1[d] | 1[d] | 2 | 2 | 0[d] | 2 |
| 8. Prospective calculation of the study size[c] | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| *Total score* | 14 | 13 | 12 | 10 | 9 | 11 | 14 | 11 | 14 | 13 | 7 | 9 | 9 | 12 | 14 | 13 | 14 |
| *Included for data extraction* | yes | yes | yes | no | no | no | yes | yes | yes | yes | no | no | no | yes | yes | no | Yes |

*a The original version contains 4 additional items for comparative studies.*
*b The items are scored 0 (not reported), 1 (reported but inadequate) or 2 (reported and adequate). The ideal total score is 16.*
*c Modified items for the purposes of our study.*
*d Studies that did not reach a full score (i.e., 2 points) on item 6 or item 7, or studies with less than 10 points on the total score were excluded for further data extraction.*

(Table 2). Studies mostly encompassed the full spectrum of common surgical techniques. Average postoperative survival periods varied from 7.7 to 14 months. Five studies mentioned postoperative complication rates ranging from 3.1 to 35%. In three studies, several other non-QoL questionnaires were completed by the included patients.[8,9,21] The included studies used a wide range of selection criteria and indications including pain, cord compression, instability, and tumor control – two studies did not describe their indication or inclusion and exclusion criteria (Appendix 3) [27,29].

### QoL Changes After Surgery

In almost all studies QoL significantly improved after 1, 3, 6, and 9 to 12 months (Appendix 4). In one study QoL decreased over the entire postoperative time.[26] Nine studies were included for the meta-analysis, as these reported means and SD's:[9,10,16,21,22,25–28] the pooled QoL summary score significantly improved from baseline to 1 month (SMD 1.09, 95% CI: 0.55, 1.64, $p < 0.001$), to 3 months (SMD 1.28, 95% CI: 0.83, 1.72, $p < 0.001$), to 6 months (SMD 1.21, 95% CI: 0.80, 1.63, $p < 0.001$), and to 9-12 months (SMD 1.08, 95% CI: 0.47, 1.70, p=0.001) (Figure 2 and Figure 3).

The $I^2$ range (92.5 to 94.6%) over all time points suggested substantial heterogeneity; however, Tau$^2$-values were $< 1$ (0.36 to 0.63). The heterogeneity may mainly be caused by publication bias –Zheng et al.[26] was the only study reporting negative effects of surgery on QoL in patients with hepatocellular carcinoma.

### Physical, Social/Family, Emotional, and Functional Well-Being

Four studies were included for FACT-G subdomain meta-analyses.[10,25,26,28]

Physical well-being improved from baseline after 1 month (SMD 0.73, 95% CI: 0.40, 1.07, $p < 0.001$), 3 months (SMD 0.94, 95% CI: 0.14, 1.75, p=0.022, but was not different at 6 months (SMD 0.52, 95% CI: -0.16, 1.19, p=0.132), nor 9-12 months (SMD -0.12, 95% CI: -1.25, 1.01, p=0.835) (Figure 4A).

Social/family well-being was not different from baseline after 1 month (SMD 0.07, 95% CI: -0.27, 0.41, p=0.675) nor 3 months (SMD 0.26, 95% CI: -0.21, 0.73, p=0.271), was improved after 6 months (SMD 0.28, 95% CI: 0.11, 0.45, p=0.001), but not different after 9-12 months (SMD 0.06, 95% CI: -0.27, 0.39, p=0.730) (Figure 4B).

Emotional well-being improved from baseline after 1 month (SMD 1.26, 95% CI: 0.44, 2.08, p=0.003), 3 months (SMD 1.27, 95% CI: 0.34, 2.20, p=0.040), 6 months (SMD 1.19, 95% CI: 0.37, 2.02, p=0.004), but was not different after 9-12 months (SMD 0.43, 95% CI: -0.62, 1.48, p=0.424) (Figure 4C).

Functional well-being improved from baseline after 1 month (SMD 1.39, 95% CI: 0.87, 1.91, $p < 0.001$), 3 months (SMD 1.34, 95% CI: 0.73, 1.95, $p < 0.001$), 6 months (SMD 1.08, 95% CI: 0.33, 1.84, p=0.005), but was not different after 9-12 months (SMD 1.44, 95% CI: -0.40, 3.27, p=0.125) (Figure 4D).

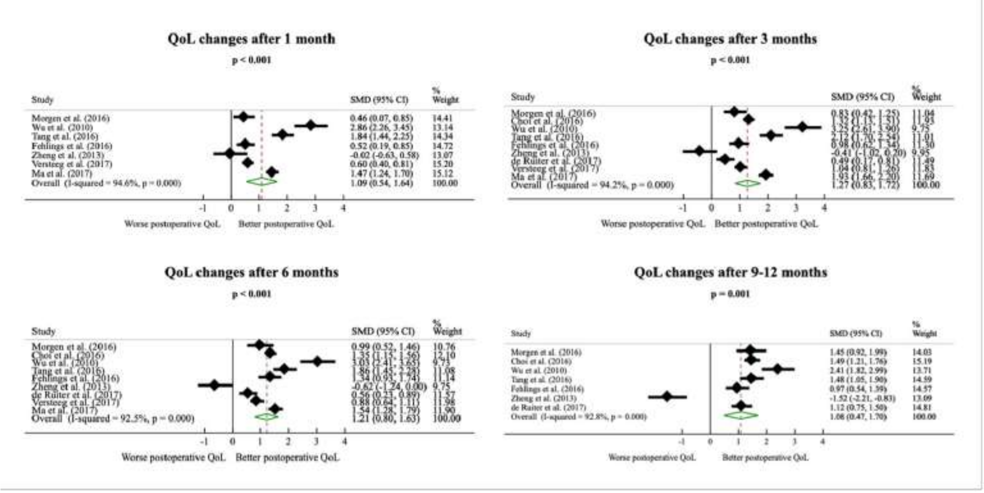


**Figure 2.** Forest plots depicting the standardized mean difference (SMD) with 95% confidence interval (CI) for quality of life (QoL) summary scores between baseline and 1 month (upper left corner), baseline and 3 months (upper right corner), baseline and 6 months (lower left corner), and baseline and 9 to 12 months (lower right corner). Five studies used the EuroQol 5 Dimensions (EQ5D) questionnaire, and four the Functional Assessment of Cancer Therapy – General (FACT-G) questionnaire. The weight per study corresponds to the weight for the meta-analysis and is based on the inverse of the variance at each time point.

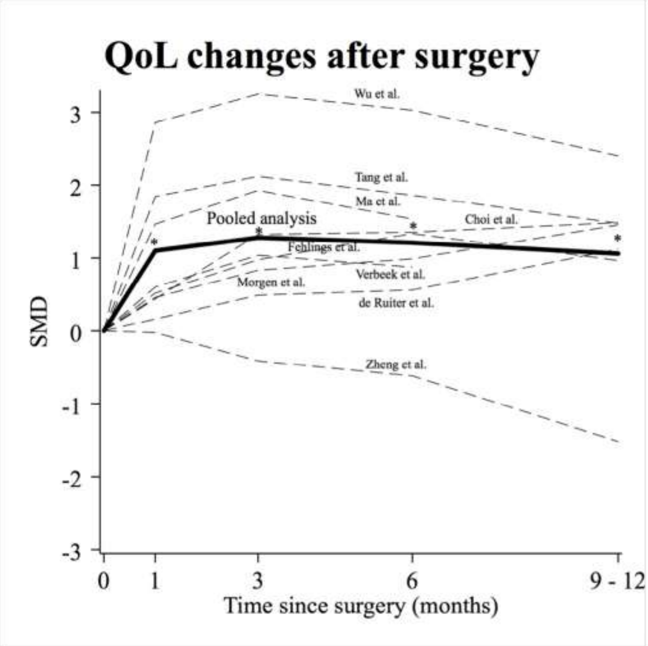
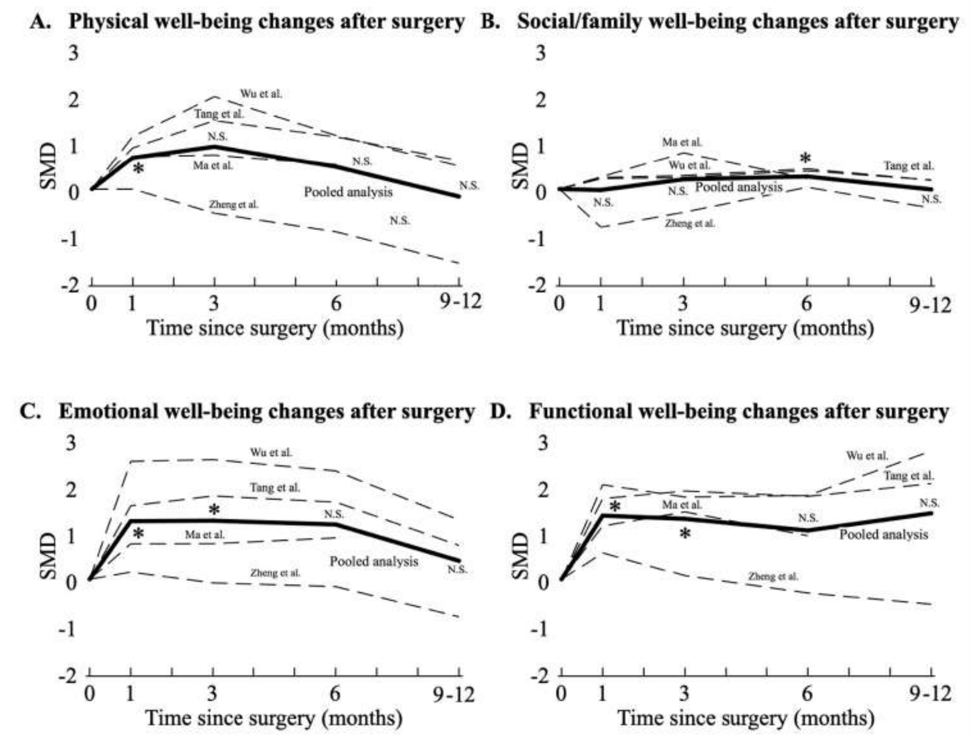


**Figure 3.** Longitudinal plot depicting changes in quality of life (QoL) after 1, 3, 6 and 9 to 12 months for studies included for this meta-analysis and for the pooled standardized mean differences (SMD). N.S.=Not Significant, *=Statistically significant ($p < 0.05$).



A. Physical well-being changes after surgery

B. Social/family well-being changes after surgery

C. Emotional well-being changes after surgery

D. Functional well-being changes after surgery

**Figure 4.** Longitudinal plot depicting changes in physical, social/family, emotional, and functional wellbeing after 1, 3, 6 and 9 months for studies included for this meta-analysis and for the pooled standardized mean differences (SMD). N.S.=Not Significant, *=Statistically significant (p < 0.05).

Overall, surgery improved compared to baseline physical well-being during the first 3-months, improved emotional and functional well-being during the first 6-months, and improved social/ family well-being at month 6.

# DISCUSSION

In the past decade, the survival of patients with spinal metastatic disease has improved substantially due to improvements in systematic therapy[30]. As a result, physicians together with patients are facing an increasing amount of management and decision making regarding the optimal treatment. However, choosing the optimal candidate for surgical intervention remains difficult. By conducting a meta-analysis, this review provides a valuable insight into the magnitude and duration of QoL benefits after surgery for patients with metastatic spinal disease which may aid this decision-making process.

This study has several limitations. First, Tang et al.[25] and Zheng et al.[26] report on specific, unfavorable, primary cancers (lung and hepatocellular) metastasizing to the spine, whereas the other eight studies included patients with various primary cancers metastasizing to the spine. The unfavorable tumor biology of the treated lesions could have led to substantially worse overall survival periods and postoperative QoL outcomes in these two studies. Second, there might have been an overlap of 7 patients with liver cancer between the study of Zheng et al.[26] (n=29) and Wu et al.[10] (n=46); this could have led to duplicate weights for these cases for the meta-analysis. However, by removing one of the two studies relevant data would be lost. Third, back pain is a major indication of surgery for metastatic spinal disease and also a crucial QoL indicator. The fact that EQ5D does, and FACT-G does not include questions regarding patients' pain level may skew the data towards more favorable QoL differences as the summary scores of the EQ5D and FACT-G questionnaires were pooled. Fourth, we included only non-percutaneous, open surgical treatments in this systematic review as we suggest these open surgical treatments to be a distinct entity within the spectrum of spinal metastasis treatment options, with different indications and postprocedural courses compared with procedures such as radiosurgery or percutaneous stabilizing techniques. The indications setting apart open surgical techniques from less invasive techniques are gross mechanical instability of the spinal column and symptomatic epidural spinal cord (or cauda equina) compression, particularly in metastases originating from radio/chemotherapy resistant tumors. In addition, patients undergoing open surgery have a different postoperative course with a longer hospitalization and rehabilitation trajectory, and an increased risk for complications[31]. Reviewing more treatment strategies and their different treatment outcomes would have been interesting but beyond the scope of this review as we focused on the effect of open surgical treatment on QoL. Future reviews could focus on the differences between treatment strategies to compare not only the indications and patient characteristics, but also QoL and other outcomes such as survival and complications.

Fifth, in five studies it was unclear how many patients were alive at each time point, making it difficult to calculate lost to follow-up percentages per time point.[10,21,25-27] Sixth, the 9 and 12 months' time points for the QoL summary scores were fused to retain statistical power. Mortality mostly

**Table 2.** Study and patient characteristics of included studies (n=10)

| Author (year) country | Study type | Surgical patients (n) | Average age, gender | Primary cancer types & survival | Surgical treatment |
|---|---|---|---|---|---|
| Morgen et al. (2016)<br>Denmark | P | 69 | Mean 64 years, 52% male | Heterogeneous<br>Median survival: 11 mo | - Posterior decompression<br>- Laminectomy<br>- Stabilization |
| Choi et al. (2016)<br>Multicenter[a] | P | 922 | Median 61 years, 56% male | Heterogeneous<br>Median survival: 14 mo | - Debulking surgery 46%<br>- Palliative decompression 36%<br>- Complete excisional surgery 18% |
| Wu et al. (2010)<br>China | P | 46 | Mean 55 years, 63% male | Heterogeneous<br>Median survival: - | - Several (remove spinal tumor, decompression, stabilization) |
| Tang et al. (2016)<br>China | P | 68 | Mean 54 years, 63% male | Lung cancer (Non-small-cell)<br>Median survival: 14 mo | - Posterior TES 10%<br>- Palliative surgery 90% |
| Fehlings et al. (2016)<br>Canada, USA | P | 142 | Mean 59 years, 58% male | Heterogeneous<br>Median survival: 8 mo | - Posterior-only approach 58.5%<br>- Anterior decompression and reconstruction 7.0%<br>- Combined anterior-posterior approach 35% |
| Zheng et al. (2013)<br>China | P | 29 | Mean 49 years, 86% male | Hepatocellular carcinoma<br>Median survival: 13 mo | - Several (excision of metastatic spinal tumors as far as possible, immediate decompression, improve neurologic function, and maintain stability of the vertebral column) |
| Versteeg et al. (2016)<br>Multicenter[b] | P | 219 | Mean 59 years, 47% male | Heterogeneous<br>Median survival: - | - Any surgical treatment |
| Ma et al. (2017)<br>China | P | 191 | mean 52 years; 59% male | Unknown primary origin<br>Median survival: - | - Circumferential decompression (54%)<br>- Laminectomy (46%) |
| Falicov et al. (2006)<br>Canada | P | 85 | Mean 59 years, 55% male | Heterogeneous<br>Mean survival: 9 mo | - Any surgical treatment that deemed fit to improve pain |
| de Ruiter et al. (2017)<br>The Netherlands | P | 94 | Mean 63 - 69 years, -% male | Heterogeneous<br>Median survival: - | - Corpectomy 50%<br>- Decompression +/- stabilization 44%<br>- Stabilization only 6.4% |

*QoL=quality of life, w=weeks, UK United Kingdom, USA=United States of America, mo=months, TES=total en bloc spondylectomy, NA=not available, UTI=urinary tract infection, PE=pulmonary embolus, DVT=deep venous thrombosis, CSF=cerebral spinal fluid, MI=myocardial infarction; a Belgium, Canada, China, Denmark, France, Japan, Netherlands, Spain, UK, USA; b Canada, Italy, Japan, Netherlands, USA; c Due to a lack of information in these studies on survival per time point, follow-up was calculated as the number of patients from that time point divided by the number of patients included at baseline (time point 0).*

| Follow-up for QoL questionnaire(s)<br>time point: patients (%) | Complications |
|---|---|
| 0w: 59 (69% from alive)<br>6w: 47 (72% from alive)<br>12w: 41 (76% from alive)<br>26w: 29 (78% from alive)<br>52w: 23 (77% from alive) | NA |
| 0mo: 332 (100% from alive)<br>3mo: 312 (99.7% from alive)<br>6mo: 209 (82% from alive)<br>12mo: 158 (67% from alive)<br>24mo: 71 (40% from alive) | **160 patients (19.6%)**<br>(Implant failure 1.4%, UTI 0.9%, Chest infection 1.7%, PE/DVT 2.0%, wound complication 3.8%, neurological 2.7%, other medical 8.7%) Surgery for recurrent tumor 2.2% |
| [c]0mo: 46 (100% from baseline)<br>1mo: 43 (93% from baseline)<br>3mo: 41 (89% from baseline)<br>6mo: 40 (87% from baseline)<br>9mo: 33 (72% from baseline) | **Death < 30 days due to complications (6.5%)**<br>(Acute liver failure 4.3%, severe pulmonary infection 2.2%) |
| [c]0mo: 68 (100% from baseline)<br>1mo: 68 (100% from baseline)<br>3mo: 68 (100% from baseline)<br>6mo: 60 (88% from baseline)<br>9mo: 45 (66% from baseline) | Overall complications: NA<br>(surgical site infections 7.4%) |
| [c]0w: 74 (100% from baseline)<br>6w: 74 (100% from baseline)<br>3mo: 59 (77% from baseline)<br>6mo: 46 (62% from baseline)<br>12mo: 34 (46% from baseline) | **42 patients (29.6%)**<br>(CSF leakage 2.1%, infection 25.0%) |
| [c]0mo: 21 (100% from baseline)<br>1mo: 21 (100% from baseline)<br>3mo: 21 (100% from baseline)<br>6mo: 21 (100% from baseline)<br>9mo: 21 (100% from baseline) | NA |
| [c]0mo: 202 (100% from baseline)<br>6w: 176 (87% from baseline)<br>12w: 148 (73% from baseline)<br>26w: 115 (61% from baseline) | NA |
| 0mo: 191 (100% from alive)<br>1mo: 178 (100% from alive)<br>2mo: 158 (100% from alive)<br>4mo: 133 (100% from alive)<br>6mo: 120 (100% from alive) | Low-grade postoperative nerve damage in **9 patients (4.7%),** without any occurrence of surgery-related complete paralysis. Postoperative infection rate in **6 patients (3.1%);** medically managed. |
| 0w: 85 (100% from alive)<br>6w: 52 (65% from alive)<br>3mo: 52 (76.5% from alive)<br>6mo: 42 (79.2% from alive)<br>12mo: 31 (79.5% from alive) | **28 patients (33%)**<br>(wound dehiscence 9.4%, CSF leak 4.7%, MI 4.7%, recurrent tumor 3.5%, instrumentation failure 3.5%, PE 2.4%, respiratory failure 2.3%, femoral nerve palsy 1.2%, nerve root injury 1.2%) |
| 0mo: 94 (100% from alive)<br>3mo: 66 (85% from alive)<br>6mo: 61 (84% from alive)<br>9mo: 51 (80% from alive)<br>12mo: 47 (81% from alive) | **33 patients (35%)**<br>Increased neurologic deficits 1.1%, dural defects 6.4%, pleural defects 1.1%, wound infection 6.4%, screw malposition/pullout 2.1%, other 19% |

depends on primary tumor type and would be expected to vary from month to month. Therefore, when interpreting the “9 to 12”-time point, it should be considered that the 3 months difference between timepoints “9 months” and “12 months” could show a clinically meaningful difference in QoL scores. Seventh, there were relatively high loss to follow-up rates for QoL measurement after surgery (up to 54% after 1 year)[21], which can partly be contributed to high mortality rates (average survival varied from 7.7-14 months) or because rapidly deteriorating patients were not able (or no longer willing) to complete questionnaires. If all patients had completed the questionnaires at all-time points, QoL scores would likely have been lower. Eight, this review may be subject to publication bias because studies that report non-significant or negative results are rarely published in surgical specialties.[32] Zheng et al.[26] was the only study reporting negative effects of surgery on QoL. Ninth, standardized validated outcome measures should be developed specifically for patients undergoing surgery for metastatic spinal disease to measure the true effect of surgery in this population. Tenth, all studies, expect Falicov et al., included patients from over 10 years ranging from 2007 to 2018. Surgical management of spinal metastases changed significantly over the last decade, including advanced minimal invasive surgical techniques which show promising results of similar pain and neurological improvement and reduced adverse events [33]. Future prospective studies should include these new techniques to investigate the QoL benefits and adverse event rates. Nevertheless, we deem the limitations proportionate to the strength of this systematic review. To our knowledge, this review provides a first overview of the magnitude and duration of QoL benefits after surgery for patients with metastatic spinal disease.

Pooled data shows that in patients operated on for various indications including pain, spinal cord compression, instability, and tumor control, QoL rapidly improved and remained stable during the first 12 months after surgery. In addition, surgery improved physical well-being during the first three postoperative months, emotional and functional well-being during the first six months, and only had effect on social/family well-being at the sixth month post-surgery.

It is often stated that the predominant goal of surgery for metastatic spinal disease is to improve QoL.[4–6] With this meta-analysis, the magnitude of these QoL changes over time were quantified. Our conclusions are based on SMD’s, but it is unclear whether these changes represented a minimal clinically important difference (MCID) –the minimal score change that reflects a meaningful change for the patient.[34] The MCID for the EQ-5D is estimated at 0.24 SMD,[35] and for the FACT-G at 0.18 SMD.[36] Using these reference values, SMD’s after 1 month (SMD 1.09), after 3 months (SMD 1.28), 6 months (SMD 1.25), and 9-12 months (SMD 1.08), however, largely surpassed MCID thresholds. Thus, our results show that the effect of surgery had a positive MCID for these patients lasting for at least 9-12 months. However, surgeons, together with patients, must weigh the likelihood of improved outcome, including pain relief, preservation of function and improved QoL, against the potential for postoperative morbidity and complications[37,38]. The improved overall QoL duration for

9-12 months relative to survival makes a surgical treatment worthwhile in at least half of the patients for one year – all studies reported at least 50% survival after 1 year. Besides postoperative morbidity, a considerable amount of patients experience complications, with rates ranging up to 35% within 30 days postoperation[39]. Severe postoperative complications include spinal epidural hematomas or leakage of cerebrospinal fluid, both with potentially devastating neurological consequences. Therefore, choosing the optimal candidate for surgical intervention in the setting of metastatic spinal disease remains crucial: otherwise, postoperative morbidity and complications may outbalance the intended benefits of surgery. In addition, the QoL improvement may not only be attributable to the operation, but also by postoperative strategies such as radiotherapy and systematic therapy.

The included studies used a wide range of indications and clinical characteristics. By pooling the studies, the results in this review cannot be applied to a specific patient population as a wide variety of indications for surgery were used, including pain, cord compression, instability, and tumor control (Appendix 3). However, stratifying the results by indication or clinical characteristics is not feasible because three studies did not provide clear indication descriptions[25,27,39] and none of the studies displayed their QoL results stratified by indication or clinical characteristics. More importantly, one never knows which component is most important to decision making. Consequently, stratifying the results into specific indications will therefore provide a false sense of accuracy as the indication is always a combination of multiple factors. Future studies should report clear definitions of selection criteria and surgical indication and provide stratified QoL results by indication and clinical characteristics, such as primary tumor type, preoperative Karnofsky and Bilsky scores, in order to elucidate the optimal candidate for surgical intervention.

The various subdomains in the FACT-G questionnaire allowed us to evaluate several important aspects of QoL among four studies:[10,25,26,28] physical, emotional, and functional well-being rapidly improved after surgery in the first three to six months, and did not drop under the baseline score after nine months. The main purpose of surgery for metastatic spinal disease is to relieve back pain and maintain/improve the ability to walk;[4–6] several items in the physical well-being (e.g. "I have pain", "I am forced to spend time in bed") and functional well-being subdomains (e.g. "I am able to work", "I am sleeping well") cover these aspects of QoL, potentially explaining improvement in these subdomains. Interestingly, emotional well-being drastically improved right after surgery; we believe that the short-term efficacy of surgery on improved clinical symptoms such as mobilization and relieve of neurological symptoms provides hope to a patient about their prognosis. Supposedly, after few months, patients encounter side-effects of (restarted) adjuvant therapies, and the cancer continues to disseminate, whereafter patients might develop awareness of a poorer prognosis, lose faith, and create fear of death –which could explain the deterioration of emotional well-being after 9-12 months[40]. Social/family well-being did not drastically change due to surgery, because patients with terminal cancer may be supported by their relatives –regardless if they are treated surgically or not[41].

Although QoL may be improved after surgery for metastatic spinal disease, it remains important to carefully select patients who would benefit from a surgical or a non-surgical treatment. Laufer et al. proposed a decision framework consisting of neurologic, oncologic, mechanical, and systemic (NOMS) considerations to determine the optimal therapy of radiation, radiosurgery, and minimally invasive and open surgical interventions[42]. If patients unfit for surgery are incautiously exposed to surgery-related morbidity or complications, QoL may not be improved at all –or even worsen[43]. Prognostication algorithms may aid in predicting outcomes such as survival, complications or clinical improvement for these patients, and thus aid in the decision-making.[44,45] However, current prognostic scoring systems in metastatic spinal disease demonstrate mixed clinical accuracy and need to be prospectively validated before they will be universally accepted by surgeons for clinical use[46,47]. Also, questionnaires that are obtained before surgery might be used as an additional tool in the shared decision making process of determining the optimal treatment.[48]

Six of the included studies compared QoL outcomes with non-surgical groups in a non-randomized manner.[10,22,25–28] Ma et al[28], Morgen et al.,[22] Wu et al.,[10], Versteeg et al.[27] and Tang et al.[25] demonstrated better QoL outcomes in the surgery group compared to the non-surgery group, indicating that surgery can improve the QoL of selected patients with metastatic spinal disease. However, carefully identifying patients that may benefit from a surgery remains important in achieving this QoL improvement. Zheng et al.[26] included patients with metastatic spinal disease of hepatocellular carcinoma; there was no difference in QoL between the surgical (n=21) and non-surgical group (n=22), and QoL for both groups declined rapidly over time. The authors explained this deterioration as the result of a lack of standard treatment for the rare spinal manifestations of hepatocellular carcinoma and the biologic aggressiveness of hepatocellular carcinoma compared to other cancer types.[49,50]

## CONCLUSION

Surgery for metastatic spinal disease rapidly improved the patients' QoL which remained stable during the first year. Rapid postoperative QoL improvement might be attributable to improved physical, emotional, and functional well-being. Our study results can be used to inform patients on postoperative expectations and help physicians understand the potential postoperative course and use this for decision-making. However, this QoL improvement may only be achieved by carefully selecting patients who may benefit from a surgery in the first place. Future research should report clear definitions of selection criteria and surgical indication and provide stratified QoL results by indication and clinical characteristics such as primary tumor type, preoperative Karnofsky and Bilsky scores to elucidate the optimal candidate for surgical intervention.

# REFERENCES

1. Mak KS, Lee LK, Mak RH, et al. **Incidence and treatment patterns in hospitalizations for malignant spinal cord compression in the United States, 1998-2006.** *Int J Radiat Oncol Biol Phys.* 2011;80(3):824–31.

2. Prasad D, Schiff D. **Malignant spinal-cord compression.** *Lancet Oncol.* 2005;6(1):15–24.

3. Cole JS, Patchell RA. **Metastatic epidural spinal cord compression.** *Lancet Neurol.* 2008;7(5):459–466.

4. Patchell RA, Tibbs PA, Regine WF, et al. **Direct decompressive surgical resection in the treatment of spinal cord compression caused by metastatic cancer: A randomised trial.** *Lancet.* 2005;366(9486):643–648.

5. Weigel B, Maghsudi M, Neumann C, et al. **Surgical management of symptomatic spinal metastases. Postoperative outcome and quality of life.** *Spine (Phila. Pa. 1976).* 1999;24(21):2240–6.

6. Kostuik J, Errico T, Gleason T, et al. **Spinal stabilization of vertebral column tumors.** *Spine (Phila. Pa. 1976).* 1988;13(3):250–256.

7. Ferrans CE. **Development of a quality of life index for patients with cancer.** *Oncol Nurs Forum.* 1990;17(3 Suppl):15–19; discussion 20.

8. Falicov A, Fisher CG, Sparkes J, et al. **Impact of surgical intervention on quality of life in patients with spinal metastases.** *Spine (Phila. Pa. 1976).* 2006;31(24):2849–2856.

9. Choi D, Fox Z, Albert T, et al. **Rapid improvements in pain and quality of life are sustained after surgery for spinal metastases in a large prospective cohort.** *Br J Neurosurg.* 2016;30(3):337–344.

10. Wu J, Zheng W, Xiao JR, et al. **Health-related quality of life in patients with spinal metastases treated with or without spinal surgery: A prospective, longitudinal study.** *Cancer.* 2010;116(16):3875–3882.

11. Ernst E, Filshie J, Hardy J. **Evidence-based complementary medicine for palliative cancer care: Does it make sense?** *Palliat Med.* 2003;17(8):704–707.

12. Choi D, Morris S, Crockard A, et al. **Assessment of quality of life after surgery for spinal metastases: position statement of the Global Spine Tumour Study Group.** *World Neurosurg.* 2013;80(6):e175-9.

13. Cheng EY. **Prospective quality of life research in bony metastatic disease.** *Clin Orthop Relat Res.* 2003;(415 Suppl):S289-97.

14. Moher D, Shamseer L, Clarke M, et al. **Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement.** *Syst Rev.* 2015;4(1):1.

15. Slim K, Nini E, Forestier D, et al. **Methodological index for non-randomized studies (minors): development and validation of a new instrument.** *ANZ J Surg.* 2003;73(9):712–6.

16. de Ruiter GCW, Nogarede CO, Wolfs JFC, et al. **Quality of life after different surgical procedures for the treatment of spinal metastases: results of a single-center prospective case series.** *Neurosurg Focus.* 2017;42(1):E17.

17. EuroQol Group. **EuroQol—a new facility for the measurement of health related quality of life.** *Health Policy (New. York).* 1990;(16):199–208.

18. Cella D, Tulsky D, Gray G. **The Functional Assessment of Cancer Therapy scale: development and validation of the general measure.** *J Clin Oncol.* 1993;11:570–579.

19. Shan L, Shan B, Suzuki A, et al. **Intermediate and long-term quality of life after total knee r eplacement. A systematic review and meta-analysis.** *J Bone Jt Surgery Am Vol.* 2015;97-A(2):156–168.

20. Shan L, Shan B, Graham D, et al. Total hip replacement: A systematic review and meta-analysis on mid-term quality of life. *Osteoarthr Cartil.* 2014;22(3):389–406.

21. Fehlings MG, Nater A, Tetreault L, et al. Survival and clinical outcomes in surgically treated patients with metastatic epidural spinal cord compression: results of the prospective multicenter AOSpine study. *J Clin Oncol.* 2016;34(3):268–276.

22. Morgen SS, Engelholm SA, Larsen CF, et al. Health-related Quality of Life in Patients with Metastatic Spinal Cord Compression. *Orthop Surg.* 2016;8(3):309–315.

23. Rücker G, Schwarzer G, Carpenter JR, et al. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol.* 2008;8:79.

24. Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *Br Med J.* 2003;327(7414):557–560.

25. Tang Y, Qu J, Wu J, et al. Effect of surgery on quality of life of patients with spinal metastasis from non-small-cell lung cancer. *J Bone Jt Surg Am Vol.* 2016;98(5):396–402.

26. Zheng W, Wu J, Xiao JR, et al. Survival and health-related quality of life in patients with spinal metastases originated from primary hepatocellular carcinoma. *J Evid Based Med.* 2013;6(2):81–89.

27. Versteeg AL, Sahgal A, Kawahara N, et al. Patient satisfaction with treatment outcomes after surgery and/or radiotherapy for spinal metastases. *Cancer.* 2019;125(23):4269–4277.

28. Ma Y, He S, Liu T, et al. Quality of life of patients with spinal metastasis from cancer of unknown primary origin: A longitudinal study of surgical management combined with postoperative radiation therapy. *J Bone Jt Surg Am Vol.* 2017;99(19):1629–1639.

29. Falicov A, Fisher CG, Sparkes J, et al. Impact of surgical intervention on quality of life in patients with spinal metastases. *Spine (Phila. Pa. 1976).* 2006;31(24):2849–2856.

30. Rosen LS, Gordon D, Tchekmedyian NS, et al. Long-term efficacy and safety of zoledronic acid in the treatment of skeletal metastases in patients with nonsmall cell lung carcinoma and other solid tumors: A randomized, phase III, double-blind, placebo-controlled trial. *Cancer.* 2004;100(12):2613–2621.

31. Barzilai O, Boriani S, Fisher CG, et al. Essential concepts for the management of metastatic spine disease: what the surgeon should know and practice. *Glob spine J.* 2019;9(1 Suppl):98S-107S.

32. Hasenboehler EA, Choudhry IK, Newman JT, et al. Bias towards publishing positive results in orthopedic and general surgery: a patient safety issue? *Patient Saf Surg.* 2007;1(1):4.

33. Pennington Z, Ahmed AK, Molina CA, et al. Minimally invasive versus conventional spine surgery for vertebral metastases: a systematic review of the evidence. *Ann Transl Med.* 2018;6(6):103.

34. Cook CE. Clinimetrics corner: the Minimal Clinically Important Change Score (MCID): a necessary pretense. *J Man Manip Ther.* 2008;16(4):82E-83E.

35. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Qual Life Res.* 2005;14(6):1523–1532.

36. King MT, Cella D, Osoba D, et al. Meta-analysis provides evidence-based interpretation guidelines for the clinical significance of mean differences for the FACT-G, a cancer-specific quality of life questionnaire. *Patient Relat Outcome Meas.* 2010:119.

37. Paulino Pereira NR, Ogink PT, Groot OQ, et al. Complications and reoperations after surgery for 647 patients with spine metastatic disease. *Spine J.* 2019;19(1).

38. Groot OQ, Ogink PT, Paulino Pereira NR, et al. **High risk of symptomatic venous thromboembolism after surgery for spine metastatic bone lesions: a retrospective study.** *Clin Orthop Relat Res.* 2019;477(7):1674–1686.

39. Fehlings MG, Nater A, Tetreault L, et al. **Survival and clinical outcomes in surgically treated patients with metastatic epidural spinal cord compression: results of the prospective multicenter AOSpine study.** *J Clin Oncol.* 2016;34(3):268–276.

40. Tong E, Deckert A, Gani N, et al. **The meaning of self-reported death anxiety in advanced cancer.** *Palliat Med.* 2016;30(8):772–779.

41. Velikova G, Booth L, Smith AB, et al. **Measuring quality of life in routine oncology practice improves communication and patient well-being: A randomized controlled trial.** *J Clin Oncol.* 2004;22(4):714–724.

42. Laufer I, Rubin DG, Lis E, et al. **The NOMS Framework: approach to the treatment of spinal metastatic tumors.** *Oncologist.* 2013;18(6):744–751.

43. Verlaan JJ, Choi D, Versteeg A, et al. **Characteristics of patients who survived 3 months or 2 years after surgery for spinal metastases: can we avoid inappropriate patient selection?** *J Clin Oncol.* 2016;34(25):3054–61.

44. Paulino Pereira NR, Janssen S, Ferrone M, et al. **Development of a prognostic survival algorithm for patients with metastatic spine disease.** *Spine J.* 2016;16(10):S318.

45. Karhade AV, Thio QCBS, Ogink PT, et al. **Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis.** *Neurosurgery.* 2019;85(1):E83–E91.

46. Schoenfeld AJ, Ferrone ML, Schwab JH, et al. **Prospective validation of a clinical prediction score for survival in patients with spinal metastases: The New England Spinal Metastasis Score.** *Spine J.* 2021;21:28-36

47. Choi D, Ricciardi F, Arts M, et al. **Prediction accuracy of common prognostic scoring systems for metastatic spine disease.** *Spine (Phila. Pa. 1976).* 2018;43(23):1678–1684.

48. van der Vliet QMJ, Paulino Pereira NR, Janssen SJ, et al. **What factors are associated with quality of life, pain interference, anxiety, and depression in patients with metastatic bone disease?** *Clin Orthop Rel Research.* 2017:475:498-507

49. Doval DC, Bhatia K, Vaid AK, et al. **Spinal cord compression secondary to bone metastases from hepatocellular carcinoma.** *World J Gastroenterol.* 2006;12(32):5247–5252.

50. Okazaki N, Yoshino M, Yoshida T, et al. **Bone metastasis in hepatocellular carcinoma.** *Cancer.* 1985;55(9):1991–1994.

# SUPPLEMENTAL MATERIAL TO CHAPTER 4

**Appendix 1.** Search syntaxes for the PubMed, Embase, and Cochrane Library on February 6th, 2020

**Appendix 2.** Quality of life questionnaires that were mentioned in the included studies for this review.

**Appendix 3.** Selection criteria, clinical characteristics and indication for surgery of included studies (n=10)

**Appendix 4.** Results of included studies (n=10)

Supplemental material can be consulted online per the website of the journal and/or publisher.

CHAPTER

05

# DO COHABITANTS RELIABLY COMPLETE QUESTIONNAIRES FOR PATIENTS IN A TERMINAL CANCER STAGE WHEN ASSESSING QUALITY OF LIFE, PAIN, DEPRESSION, AND ANXIETY?

Olivier Q. Groot, Nuno R. Paulino Pereira, Michiel E.R. Bongers, Paul T. Ogink, Erik T. Newman, Jorrit-Jan Verlaan, Kevin A. Raskin, Santiago A. Lozano-Calderon, Joseph H. Schwab

*Clinical Orthopaedics and Related Research 2021;479:792-801.*

# ABSTRACT

### Background

Patients with bone metastases often are unable to complete quality of life (QoL) questionnaires, and cohabitants (such as spouses, domestic partners, offspring older than 18 years, or other people who live with the patient) could be a reliable alternative. However, the extent of reliability in this complicated patient population remains undefined, and the influence of the cohabitant's condition on their assessment of the patient's QoL is unknown.

### Objectives

(1) Do QoL scores, measured by the 5-level EuroQol-5D (EQ-5D-5L) version and the Patientreported Outcomes Measurement Information System (PROMIS) version 1.0 in three domains (anxiety, pain interference, and depression), reported by patients differ markedly from scores as assessed by their cohabitants?

(2) Do cohabitants' PROMIS-Depression scores correlate with differences in measured QoL results?

### Design

Cross-sectional survey study.

### Methods

This study included patient and cohabitant age older than 18 years; presence of histologically confirmed bone metastases (including lymphoma and multiple myeloma) and presence of the cohabitant at the clinic visit. Patients were eligible for inclusion in the study regardless of comorbidities, prognosis, prior surgery, or current treatment. Between June 1, 2016 and March 1, 2017 and between October 1, 2017 and February 26, 2018, all 96 eligible patients were approached of whom 49% (47) met the selection criteria and were willing to participate. The outcomes measures were 5-level EuroQol-5D and the Patient Reported Outcomes Measurement Information System (PROMIS) version 1.0 in three domains (anxiety, pain interference, and depression).

### Results

There were no clinically important differences between the scores of patients and their cohabitants for all questionnaires, and the agreement between patient and cohabitant scores was moderate to strong (Spearman's correlation coefficients ranging from 0.52 to 0.72 on the four questionnaires: all p values < 0.05). However, despite the good agreement in QoL scores, an increased cohabitant's depression score was correlated with an overestimation of the patient's symptom burden for the anxiety and depression domains (weak Spearman's correlation coefficient (95% confidence interval), of 0.33 [95% CI 0.08 to 0.58]; p=0.01 and moderate Spearman's correlation coefficient of 0.52 [95% CI 0.29 to 0.74]; p < 0.01, respectively).

**Conclusion**

The present findings support that cohabitants might be reliable raters of the QoL of patients with bone metastases. However, if a patient's cohabitant has depression, the cohabitant may overestimate a patient's symptoms in emotional domains such as anxiety and depression, warranting further research that includes cohabitants with and without depression to elucidate the effect of depression on the level of agreement. For now, clinicians may want to reconsider using the cohabitant's judgement if depression is suspected.

# INTRODUCTION

Patients with metastatic bone disease have relatively short life expectancies, and their disease often leads to substantial pain, disability, and decreased quality of life (QoL).[1] The main aim of treatment is to reduce pain and restore function.[2,3] This makes an assessment of patient-reported outcomes important in understanding and quantifying the effectiveness of treatment on the patients' perceptions about their health.[4] Patients with advanced disease may find questionnaire completion to be physically or emotionally burdensome, making self-reported QoL instruments less-feasible and accurate.[5] Therefore, it is valuable to know the validity and reliability of patient-related outcomes provided by cohabitants (such as spouses, domestic partners, offspring older than 18 years, or other people who live with the patient).

Physical QoL scores reported by patients with metastatic disease and those reported by their spouses are more concordant than those reported by patients and their physicians[6], but studies have had conflicting results about the extent of agreement. Among patients with metastatic prostate carcinoma, a high accuracy in spousal evaluation of physical and psychosocial functioning, symptoms, and overall QoL was reported.[5] However, a comparable study in patients with metastatic prostate or breast cancer demonstrated the opposite, with substantial variability in QoL scores between physicians and partners and patients.[7] Perception of quality of life by patients, partners and treating physicians. To our knowledge, no such studies have been performed in patients with metastatic bone disease and their cohabitants. In addition, a high prevalence of depression has been found in people who live with patients with cancer[8,9]; no prior studies that we know of have assessed whether the cohabitant's mental and emotional condition influences his or her capability of judging the patient's QoL.

We therefore asked: (1) Do QoL scores, measured by the EuroQol-5D 5-level (EQ-5D-5L) version and the Patient-reported Outcomes Measurement Information System (PROMIS) version 1.0 in three domains (anxiety, pain interference, and depression), reported by patients differ markedly from scores as assessed by their cohabitants? (2) Do cohabitants' PROMIS-Depression scores correlate with differences in measured QoL results?

# METHODS

### Study Design and Setting

This cross-sectional study design was approved by the institutional review board of our tertiary care hospital. Patients attending orthopaedic oncology visits with one of three orthopaedic oncologists (JHS, KAR, SALC) at our institution were approached for inclusion in the study between the two periods, June 1, 2016 and March 1, 2017 and between October 1, 2017 and February 26, 2018; the gap was caused because of the departure of a clinical research assistant and the need to hire a new one. All three enrolling orthopaedic oncologists were active during these two periods and patient care did not change.

### Participants

The inclusion criteria were patient and cohabitant age older than 18 years; presence of histologically confirmed bone metastases, including lymphoma and multiple myeloma; presence of the cohabitant at the clinic visit, defined as a person with whom the patient shares his or her living space (for example, spouse, offspring older than 18 years, or close friend), and patients and cohabitants who were proficient in English. The presence of a cognitive impairment in either the patients or their cohabitants that could limit questionnaire completion, as judged by the treating attending physician, was an exclusion criterion. Patients were eligible for inclusion in the study regardless of their comorbidities, prognosis, prior surgery, or current treatment. Eligible patients and cohabitants received a full verbal and written explanation of the purpose and procedures of the study.

### Demographics, Description of Study Population

Ninety-six patients were approached, including 35 who were not accompanied by a cohabitant, eight who were not proficient in English, four who refused to participate, and two who were unable to complete the questionnaires because of their health status (Appendix 1). Thus, 47 patients with bone metastases were enrolled. Additional analyses of baseline and disease characteristics between the included (n=47) and excluded (n=49) groups demonstrated no differences between them, including the Katagiri et al.[10] survival score and modified Charlson comorbidity index score[11] (Appendix 2). The median patient age was 69 years (interquartile range 63 to 74 years) (Table 1). Patients had a wide range of cancer diagnoses, with breast cancer (21%; 10 of 47 patients), kidney cancer (17%; 8 of 47

**Table 1.** Sociodemographic and clinical characteristics (n=47)

| *Variables* | *n=47* |
|---|---|
| Median age of patient in years (IQR) | 69 (63 to 74) |
| Median years living together (IQR) | 41 (27 to 49) |
| Median Modified Charlson Comorbidity Index (IQR) | 6 (6 to 6) |
| Duration in years of primary diagnosis until enrollment (IQR) | 5 (2 to 11) |
| Men, % (n) | 43 (20) |
| Education of patient as terminal degree, % (n) | |
| High school or below | 30 (14) |
| College/bachelor degree | 32 (15) |
| Graduate/professional degree | 38 (18) |
| Education of cohabitant as terminal degree, % (n) | |
| High school or below | 23 (11) |
| College/bachelor degree | 49 (23) |
| Graduate/professional degree | 28 (13) |
| Cohabitant relation, % (n) | |
| Spouse | 74 (35) |
| Child | 13 (6) |
| Domestic partner | 6 (3) |
| Non-domestic partner | 4 (2) |
| In a relation | 2 (1) |
| ECOG performance status[a], % (n) | |
| Good score 0-2 | 96 (45) |
| Poor score 3-4 | 4 (2) |
| Primary cancer, % (n) | |
| Breast cancer | 21 (10) |
| Kidney cancer | 17 (8) |
| Sarcoma | 13 (6) |
| Lung cancer | 11 (5) |
| Prostate cancer | 9 (4) |
| Other[b] | 30 (14) |
| Location of histologically confirmed bone metastasis, % (n) | |
| Femur | 34 (16) |
| Pelvis | 28 (13) |
| Spine | 26 (12) |
| Thoracic spine | 17 (8) |
| Lumbar spine | 9 (4) |
| Cervical spine | 0 (0) |
| Other[c] | 13 (6) |
| Visceral metastases[d], % (n) | |
| Yes | 36 (17) |
| No | 64 (30) |
| Multiple bone metastases, % (n) | |
| Yes | 66 (31) |
| No | 34 (16) |
| Prior surgery impending/pathologic fracture (within 1 year), % (n) | |
| Yes | 23 (11) |
| No | 77 (36) |

*Continued on next page*

| | |
|---|---|
| Ethnicity, % (n) | |
| Caucasian | 98 (46) |
| Hispanic | 2 (1) |
| Currently using pain medication[e], % (n) | |
| Yes | 64 (30) |
| No | 28 (13) |
| Radiotherapy for the bone lesion, % (n) | |
| Yes, currently | 11 (5) |
| Yes, in the past | 38 (18) |
| No | 51 (24) |
| Chemotherapy[e], % (n) | |
| Yes, currently | 26 (12) |
| Yes, in the past | 32 (15) |
| No | 38 (18) |
| Other disabling conditions[e,f], % (n) | |
| Yes | 30 (14) |
| No | 68 (32) |
| Future surgery impending/pathologic fracture (within 3 months), % (n) | |
| Yes | 17 (8) |
| No | 83 (39) |

*IQR=interquartile range; ECOG=Eastern Cooperative Oncology Group.*
*a The ECOG performance status was dichotomized into good scores 0-2 (50% of waking hours bed or chair bound) or poor scores 3-4 (> 50% of waking hours bed or chair bound).*
*b This category includes multiple myeloma 4.3% (2), lymphoma 4.3% (2), skin cancer 4.3% (2), colorectal cancer 2.1% (1), thyroid cancer 2.1% (1), esophageal cancer 2.1% (1), melanoma 2.1% (1), giant cell tumor 2.1% (1), cholangiocarcinoma 2.1% (1), ovarian 2.1% (1) and unknown 2.1% (1).*
*c This category includes rib 4.3% (2), humerus 4.3% (2), sacrum 2.1% (1), and tibia 2.1% (1)*
*d This category includes metastases to lung, liver, and/or brain.*
*e Missing data in patients were currently using pain medication 9% (4); chemotherapy 4% (2); and other disabling conditions 2% (1).*
*f This category consists of a self-reported measure by the patient in the questionnaire, posed as "Do you have any other disabling conditions?"*

patients), and sarcoma (13%; 6 of 47 patients) being the most prevalent. Twenty-three percent (11 of 47 patients) underwent surgery for an impending or complete pathological fracture surgery before enrollment, and 77% (36 of 47 patients) did not receive any surgical treatment for their bone lesion. Seventeen percent (8 of 47 patients) were in their preoperative period because they received surgery within 3 months of enrollment for an impending or complete pathological fracture of a bone lesion. The median cohabitant age was 41 years (IQR 27 to 49 years); most (74%; 35 of 47 patients) of the cohabitants were spouses.

### Description of Experiment

The patient and cohabitant were offered a tablet on which to complete the questionnaires simultaneously; the patient and cohabitant were asked to complete it without discussing their answers with each other. A researcher (OQG, NRPP, PTO) was present in the room to provide instructions and ensure independent completion of the questionnaires. We did not use any modified versions of the surveys, but we gave clear standard instructions to the cohabitant to complete the surveys from the patient's perspective. The questionnaires were automatically saved and processed anonymously, which meant the researcher could not ensure completion of all questions.

### Outcomes and Explanatory Variables

QoL was assessed with the EQ-5D-5L[12] and PROMIS version 1.0 short forms for pain interference, anxiety, and depression.[13] Both surveys are self-administered and broadly used in clinical practice to measure a variety of QoL domains. Patients and cohabitants received the same four QoL questionnaires: the EQ-5D-5L, PROMIS-Pain Interference 8a, PROMIS-Anxiety 8a, and PROMIS-Depression 8a. In addition, the cohabitant received a fifth QoL questionnaire that evaluated the cohabitant's depression status (PROMIS-Depression 4a). Regarding the two PROMIS-Depression questionnaires, the first had eight questions to assess the patient's depression score and the second had four questions to assess their depression score. Additional non-QoL-related questions asked about education as terminal degree (high school or below, some college education or bachelor's degree, or graduate or professional degree, by the patient and cohabitant), total years of cohabitation (by the patient), the current use of pain medication (yes or no, by the patient), and the presence of other disabling conditions (yes or no, by the patient).

The EQ-5D-5L is composed of one question for each of five items (mobility, self-care, usual activities, pain or discomfort, and anxiety or depression) through five possible answers on a five-point Likert scale ranging from 1 ("no problems with") to 5 ("major problems with")[14]. The combination of answers is converted into a single EQ-5D-5L score ranging from 0 (death) to 1 (perfect health). The EQ-5D-5L also includes a VAS, in which respondents can rate their perceived health status from 0 ("worst health you can imagine") to 100 ("best health you can imagine"). A higher score on the EQ5D-index and EQ-VAS indicates a better health status.

The PROMIS questionnaires are composed of eight questions for each of the three domains (pain interference, anxiety, and depression) through five possible answers on a five-point Likert scale ranging from 1 ("no problems with") to 5 ("major problems with"). The combination of eight answers for each domain is converted into a T-score metric that is normalized with respect to the general US population (mean=50; SD=10).[15] A higher T-score represents more of the domain being reported. For example, a mean PROMIS-Anxiety T-score of 60 indicates an increased anxiety level of one SD above the general population mean. Anchor-based MCIDs for patients with various advanced-stage cancer (Stages III and IV) receiving any type of treatment were available for all measured questionnaires: EQ-5D 0.10, VAS 12, pain interference 6.1, anxiety 4.6, and depression 4.3.[16,17] The upper boundary of the estimates was considered since patients with bone metastases are a complicated patient population.

The following explanatory variables were collected by a researcher (MERB) blinded for the outcome to assess any influence on the level of agreement: age, modified Charlson Comorbidity Index[11], Eastern Cooperative Oncology Group (ECOG) performance status[18], primary tumor type, Katagiri survival score[10], duration of primary diagnosis until enrollment, location of histology confirmed bone metastasis, visceral metastases (lung, liver, and/or brain), multiple bone metastases, prior surgery for impending/pathologic fracture (within 1 year of enrollment), radiotherapy for bone lesion, and future surgery for impending/pathologic fracture (within 3 months after enrollment).

### Accounting for all Patients

There were missing data for 36 questions distributed over nine patients and 12 cohabitants. The largest number of unanswered questions by patients was three and two for cohabitants. No questionnaire had more than one unanswered question. The questionnaires with missing data were included because the converted EQ-5 index scores and PROMIS T-scores were corrected, with the response pattern scoring for a maximum of one and four missing responses, respectively. An additional worst-case analysis demonstrated no changes in the results.

### Statistical Analysis

Descriptive statistics were used to characterize the study population. The median and IQR were calculated because the score distributions were non-normally distributed, with a skew towards lower EQ-VAS scores and EQ-5D -5L scores and higher PROMIS T-scores. To evaluate baseline differences between the included and excluded groups, we used a t-test for continuous variables and Fisher's exact test for categorical variables.

Patient-cohabitant agreement was assessed with three methods, in accordance with previously reported approaches, enabling a comparison of results among studies.[5-7,19,20] First, Wilcoxon's signed-rank test was used to assess differences in the EQ-5D-5L, PROMIS-Pain Interference, PROMIS-Anxiety, and PROMIS-Depression scores between patients and cohabitants (individually paired groups). Spearman's rank correlation was used to determine the relationship between the individual patient and cohabitant scores, as follows: 0.00 to 0.19, very weak agreement; 0.20 to 0.39, weak agreement; 0.40 to 0.59, moderate agreement; 0.60 to 0.79, strong agreement; and 0.81 to 1.00 excellent agreement. This correlation increases in magnitude toward 1 as patient and cohabitant scores become closer to being perfectly in line with each other. A negative correlation would indicate an inverse association; that is, patient scores decreasing as cohabitant scores increase, or the other way around. A correlation coefficient of zero indicates that no relationship exists between the scores of patient and cohabitant. We used bootstrapping (number of resamples: 1000) to calculate p values and 95% CIs for the Spearman's rank correlation coefficients.

Second, to compare the overall level of agreement[5,6], we calculated the percentage agreement of all individual questions from the EQ-5D-5L and the three PROMIS domains, using the actual 5-point response format. Thus, 29 items per patient-cohabitant pair (five from the EQ-5D-5L and eight from each of the PROMIS domains) were assessed for exact agreement or a 1- to 4-point response difference to show the extent of agreement of the individual items. The 0- to 100-point score of the EQ-VAS was excluded from this analysis to enable unambiguous interpretation. A total of 1363 comparisons were possible from the 29 items assessing QoL on a 5-point scale for each of the 47 patient-cohabitant pairs. However, 1329 comparisons between patient and cohabitants were made because 2.5% (34 of 1363) of the questions had missing answers.

Third, to investigate disagreement between patients and cohabitants[6], we calculated the percentage of cohabitants who underestimated or overestimated the patient's symptom burden on all individual questions. The same test to measure patient-cohabitant agreement, the Spearman's rank correlation, was used to assess whether underlying depression in the cohabitant, self-reported using the shorter PROMIS-Depression (4a), correlated with differences in questionnaire scores as reported by both patients and their cohabitants. A positive coefficient indicates that an increased cohabitant's depression score correlates with an overestimation of the patient's symptoms. A negative coefficient corresponds with an inverse association: An increased cohabitant's depression score correlates with an underestimation of the patient's symptoms.

The minimally clinically important difference for the PROMIS questionnaires in a comparable population of patients with spinal metastases is 4.5, with an SD of 10.[1] A two-tailed paired t-test demonstrated that to find a difference of 4.5 between the patient and cohabitant responses (effect size 0.45), we needed at least 47 patients (alpha 0.05; power 0.85). A two-tailed P-value < 0.05 was considered significant. Prospective enrollment continued until this target was reached. Patients were included only once to respect the statistical rule of independence. The questionnaires were administered by Assessment Center and REDCap.[13,21] All statistical analyses were performed using Stata version 13.0 (StataCorp LP, College Station, TX, USA).

# RESULTS

### Differences Between Patients and Cohabitants in QoL Scores

There were no clinically important differences between the scores of patients and their cohabitants for all questionnaires, and the agreement between patient and cohabitant scores was moderate to strong (Spearman's correlation coefficients ranging from 0.52 to 0.72 on the four questionnaires: all P-values < 0.05, Table 2). We determined there was exact agreement in 45% (607 of 1329) of the correlations, disagreement by 1 point out of 4 in 39% (525 of 1329), and disagreement of 2 or more points out of 4 in 15% (197 of 1329) (Table 3). The disagreements were generally because the cohabitant overestimated the symptom burden on all questionnaires (Table 4).

**Table 2.** Comparison of patient and cohabitant scores for the completed questionnaires.

| | *Median (interquartile range)* | | *P-value*[a] | *Spearman correlation coefficient (95% confidence interval)*[b] |
|---|---|---|---|---|
| | *Patient* | *Cohabitant* | | |
| EQ-VAS* | 70 (50-80) | 70 (50-80) | 0.34 | 0.52 (0.26-0.77) |
| EQ-5D index values* | 0.73 (0.56-0.83) | 0.70 (0.54-0.78) | 0.59 | 0.72 (0.56-0.88) |
| PROMIS Pain Interference^ | 63 (52-66) | 62 (56-67) | 0.18 | 0.69 (0.56-0.83) |
| PROMIS Anxiety^ | 54 (47-60) | 54 (49-62) | 0.10 | 0.66 (0.45-0.87) |
| PROMIS Depression^ | 50 (38-57) | 52 (38-60) | 0.35 | 0.56 (0.36-0.76) |

*EQ-VAS=EuroQol Visual Analogue Scale, EQ-5D=EuroQol 5 dimensions, PROMIS=Patient-Reported Outcomes Measurement Information System,*
*a P-value calculated by Wilcoxon signed rank*
*b Spearman rank correlation with 95% confidence interval calculated through bootstrapping (1,000 resamples). Each coefficient has a p<0.001.*
*c Higher score represents better health status; P-values remain not significant after worse case analysis for missing data in EQ-VAS (incomplete dyads=2 (4.3%); p=0.215) and EQ-5D index values had at least 4 items scored within each questionnaire and hence corrected with the response pattern scoring.*
*d Higher score represents a greater degree of symptoms in the quality of life; at least 4 items were scored within each questionnaire and hence corrected with the response pattern scoring.*

**Table 3.** Comparison of patient and cohabitant scores for the EQ-5D-5L (n=235) and three PROMIS questionnaires (n=376 each questionnaire) for a total of 1363 comparisons on a 5-point response scale.

| *QoL questionnaire* | *Agreement, % (n)* | *1-point difference, % (n)* | *2-point difference, % (n)* | *3-point difference, % (n)* | *4-point difference, % (n)* | *Missing data, % (n)* | *Total, n* |
|---|---|---|---|---|---|---|---|
| EQ-5D-5L | 46 (108) | 41 (96) | 8.9 (21) | 1.3 (3) | 0 (0) | 3.0 (7) | 235 |
| PROMIS Pain Interference | 40 (151) | 37 (148) | 18 (66) | 2.1 (8) | 0.3 (1) | 3.2 (12) | 376 |
| PROMIS Anxiety | 41 (154) | 43 (161) | 11 (42) | 2.4 (9) | 0.3 (1) | 2.4 (9) | 376 |
| PROMIS Depression | 52 (194) | 35 (130) | 10 (39) | 1.9 (7) | 0 (0) | 1.6 (6) | 376 |
| Totals | 45 (607) | 39 (525) | 12 (168) | 2.0 (27) | 0.1 (2) | 2.5 (34) | 1363* |

*EQ-5D-5L=EuroQol 5 dimensions 5 levels (5 questions scored on a 5-point scale); PROMIS=Patient-Reported Outcomes Measurement Information System (every PROMIS domain consists of 8 questions scored on a 5-point scale)*
**Across five questions for EQ-5D-5L for 47 patients (5 x 47=235 comparisons) and eight questions for each of the three PROMIS questionnaires (8 x 47=376 comparisons for one PROMIS questionnaire); in total 235 (EQ-5D-5L) + 3 x 376 (PROMIS)=1363 comparisons of which 2.5% (34) are missing; all questionnaires have a 5-point response format, excluding the EQ-VAS which uses the 0-100 scale; agreement represents exact agreement on the 5-point response scale, 1-point difference represents disagreement by 1-point, and so on.*

**Table 4.** Correlation between underlying cohabitants' depression with differences in completed questionnaires

| | *Cohabitant* | | | | |
|---|---|---|---|---|---|
| | *Agreement, n (%)* | *Overestimates, n (%)* | *Underestimates, n (%)* | *Spearman correlation coefficient (95% CI)*[a] | *P-value* |
| EQ-VAS[b, c] | 21 (45) | 10 (21) | 14 (30) | 0.20 (-0.12 to 0.52) | 0.22 |
| EQ-5D-5L[b] | 108 (46) | 40 (17) | 80 (34) | 0.13 (-0.19 to 0.45) | 0.43 |
| PROMIS Pain Interference[b] | 151 (40) | 80 (21) | 133 (35) | 0.01 (-0.27 to 0.31) | 0.92 |
| PROMIS Anxiety[b] | 154 (41) | 79 (21) | 134 (36) | 0.33 (0.08 to 0.58) | **0.01** |
| PROMIS Depression[b] | 194 (52) | 63 (17) | 113 (30) | 0.52 (0.29 to 0.74) | **< 0.01** |

*EQ-VAS=EuroQol VAS; EQ-5D-5L=the five-level EuroQol-5D; PROMIS=Patient-Reported Outcomes Measurement Information System.*
*a Spearman's rank correlation with 95% confidence interval calculated through bootstrapping (1000 resamples); an increased correlation coefficient that reflects the underlying cohabitants' depression increases the difference between the survey outcome, with the cohabitant estimating the patient's health more pessimistically compared with the patient's own rating.*
*b Missing data were for EQ-VAS 4.3% (2), EQ-5D-5L 3.0% (7), PROMIS Depression 1.6% (6), PROMIS Anxiety 2.4% (9), and PROMIS Pain Interference 3.2% (12); additional worse case analyses demonstrated no significant changes in p values;* **bold** *indicates significance (P<0.05) based on Spearman's rank correlation comparing the cohabitants' depression and the differences in questionnaires outcomes completed by the patients and their respective cohabitant; VAS scale, EQ-5D-index and PROMIS T-scores were used in this analysis.*
*c Level of agreement was based on a ±10-point range on the 0-100 scale.*

### Correlation of Cohabitants' PROMIS-Depression Scores with Patients' QoL Results

The cohabitant's overestimation of the symptom burden for the anxiety and depression domains was associated with the cohabitant's depression score (Spearman's correlation coefficient 0.33 [95% CI 0.08 to 0.58]; p=0.01 and Spearman's correlation coefficient 0.52 [95% CI 0.29 to 0.74]; p<0.01, respectively). The correlation is positive, which means that a cohabitant's increased depression score correlates with an overestimation of the patient's symptoms. This degree of correlation was weak for anxiety and moderate for depression. The observed overestimation of symptoms on the EQ-VAS, EQ-5D-5L, and PROMIS-Pain was not associated with the cohabitant's depression score.

# DISCUSSION

The assessment of QoL plays an increasingly important role in patients with bone metastases. Although patients prefer QoL assessments[22], 40% to 70% of patients who are critically ill are unable or unwilling to complete QoL questionnaires.[19] Cohabitants could play a major role as a reliable alternative. The use of alternative raters has been investigated, but prior studies were either not specifically designed for patients with bone metastases or did not investigate the use of only cohabitants as raters for patients with bone metastases, and the results are inconsistent.[6,7,19,20,23–25] Moreover, to our knowledge, no previous study assessed the influence of the cohabitant's depression status on his or her capability of judging the patient's QoL. Our goals in this study were to assess differences between patient- and cohabitant-perceived QoL, pain, depression, and anxiety, and to assess whether the cohabitants' depression scores correlated with differences in measured QoL results. For all QoL questionnaires in the present study, there was moderate-to-strong agreement between patients and their cohabitants. However, despite the good agreement in QoL, the cohabitants' higher depression scores were correlated with increased differences in the anxiety and depression domains on the PROMIS. These findings are important because cohabitant QoL scores could be used to evaluate patients with bone metastases, although the cohabitant's depression may cause overestimation of the patient's symptoms.

This study has limitations. First, the observed results may only apply to patients who are able to complete a questionnaire. Our findings might be extrapolated to patients with impairments through future multi-institutional studies with larger numbers, and a generic function for patient-cohabitant agreement in QoL could be developed. Second, selection bias may have occurred. We included patients and cohabitants who arrived at the clinic together and were both willing to participate, possibly indicating a more emphatic and intimate relationship. This potentially biased our findings toward a higher degree of agreement between patients and their cohabitants. Additionally, half of the approached patients were excluded. However, the included and excluded groups did not differ in baseline or disease characteristics. Furthermore, we only included patients with bone metastases who were seen in an orthopaedic office, most likely representing a frailer population. Third, enrollment did not occur during a 7-month period because of a delay in clinical research employment. However, we believe that this did not influence our results or the randomness of our sample. No changes in patient care or orthopaedists were noted during this period. Fourth, the observed differences between the scores are partly explained by the cohabitants' depression scores. Additional analyses of differences with respect to the duration of cohabitation, education of the patient or cohabitant, and cohabitant relationship to the patient demonstrated no relationship. We may have overlooked other specific variables that could have influenced differences in scores, such as economic stability, domestic health, and overall QoL of the cohabitant. Future research should study these factors to clarify the relationship between patient and cohabitant scores. Fifth, only a single timepoint in the disease process was measured, and the level of agreement may fluctuate over time. For example, a patient recently diagnosed with cancer may not have had time with his or her cohabitant to adjust to the situation thus potentially resulting in disparate scores whereas the cohabitant and patient with a long-standing history of cancer have acclimated to the situation might be better aligned in the scoring. Although additional analyses among disease characteristics such as duration of primary diagnosis until enrollment, surgery within 3 months of enrollment, or prior surgery had no influence on the level of agreement, stage of disease and time may be a factor in equilibrating scores between patient and cohabitant. However, this study is underpowered with respect to this relationship and was not designed to investigate it. Future study designs should incorporate this time element in their protocol and investigate its influence on QoL agreement by including patient-cohabitant pairs at different time points in their disease process, such as recently diagnosed bone metastases, undergoing active chemotherapy, or those in the perioperative period.

The scores for all studied QoL domains were concordant between patients with bone metastases and their cohabitants, with moderate-to-strong Spearman's correlation coefficients. The range of proportions of exact agreement in our study were generally better than those reported in previous studies of patients with cancer: 40% to 52% and 18% to 68% for patients and their cohabitants, respectively.[7,19,20] A possible reason for our high correlation and agreement compared with those in previous studies is that these studies included family members, family caregivers, and proxies (defined as "persons in close relationship with the patient")[7,19,20], compared with cohabitants consisting largely of spouses. When evaluating the 1329 comparisons for the individual questions, we found that more than half disagreed (54%; 722 of 1329). However, most disagreements (73%; 525 of 722) fell within one response category (for example, "slight problems with" versus "moderate problems with"). This is represented by the good Spearman's correlation coefficients. Most disagreements (66%; 474 of 722) were overestimations of the patient's symptom burden by the cohabitant; this has also been demonstrated in multiple studies.[7,20,26–28] The overestimation may be explained by the response-shift bias: a shift in the frame of reference for scoring overall QoL [28]. This systematic bias, however, was generally small. The results, in concordance with those of previous studies[5,20,24,25], support the use of cohabitants by clinicians in circumstances where the patient is unable to complete a questionnaire because of cognitive impairment, communication deficits, serious emotional or physical distress, a language barrier, or unwillingness. In addition, increasing the cohabitant's engagement in the patient's disease progress is desirable to ensure appropriate care, sense of involvement, and improvement of shared decision-making. However, further research is required to investigate the influences of situational changes in the disease process on the level of agreement by employing longitudinal study designs.

Research designs can benefit too from the use of cohabitants to determine patients' QoL scores. First, longitudinal studies that include QoL endpoints can be impaired due to missing QoL data, especially in patient populations with advanced disease, such as bone metastases, in whom severe symptomology or disease progression impedes patients from completing QoL questionnaires. Using cohabitants' scores can enhance the quality of such studies by decreasing the amount of missing QoL data. Second, cross-sectional studies might perform a more representative evaluation of QoL by including cohabitants' scores[29], although it is advisable to obtain a substantial portion of scores from both the patient and cohabitant to ensure that the cohabitant's assessment of QoL is accurate.[20] Lastly, clinical trials could include more completed QoL questionnaires, resulting in less-biased comparisons between treatments.[30]

The cohabitant's mental and emotional condition is adversely affected because she or he must fulfill a demanding role in managing a patient's malignant disease and supporting them.[31–33] Unsurprisingly, a high prevalence of depression has been found in people who live with patients who have cancer.[8,9] We thought that the cohabitant's depression would affect his or her capability to judge the patient's

QoL, and this was partially supported. Underlying depression in the cohabitant is associated with increased disagreement in the emotional domains (anxiety and depression). However, cohabitants can empathize with patients, based on QoL domains that involve less emotional judgement such as pain, VAS perceived health, and EQ-5D-5L (only one of the five questions are an emotional domain), regardless of their depression status. The cohabitant's depression score should be included in the assessment of the patient's QoL and the scores should be corrected for to prevent a potential overestimation of the symptom burden. Future research is required to elucidate an accurate correction by comparing two cohabitants group, with and without depression, and the effect of depression on the level of agreement. For now, clinicians may want to reconsider using cohabitants' judgements when a cohabitant shows signs of depression.

Some have suggested that treating physicians are in a good position to rate a patient's QoL, but multiple studies indicated that partners or other close relatives are a more-reliable alternative.[7,20,29,34] Proxies, unlike treating physicians, observe patients during an extended period in a range of circumstances and are less biased toward a course of treatment that often depends on the QoL score. Additionally, physicians vary considerable during treatment. As such, and based on our findings in this study, we strongly recommend the use of cohabitants rather than treating physicians when an individual patient cannot complete QoL scores and evaluating that patient's QoL is important as part of that patient's care.

## CONCLUSION

Our findings indicate that for patients with bone metastases, cohabitants may be a reliable alternative to patients who are unable to complete QoL questionnaires, although patient self-reported QoL is preferred. However, clinicians may want to reconsider relying on cohabitants' judgements if they show signs of depression. Refining the association between patient and cohabitant QoL scores requires further research. We believe that employing a range of questionnaires that investigate the cohabitant's own health status might provide further insight into how the cohabitant's QoL affects his or her judgement. Second, the influence of situational changes in the disease process on the level of agreement should be investigated; specifically, we would recommend longitudinal study designs and deeper inquiry into the possible influence of a new diagnosis of metastatic disease, particular chemotherapeutic regimes (or active chemotherapy more generally), and patients who are in the perioperative period.

# REFERENCES

1. Choi D, Fox Z, Albert T, et al. **Rapid improvements in pain and quality of life are sustained after surgery for spinal metastases in a large prospective cohort.** *Br J Neurosurg.* 2016;30(3):337–44.

2. Steensma M, Healey JH. **Trends in the surgical treatment of pathologic proximal femur fractures among musculoskeletal tumor society members.** *Clin Orthop Relat Res.* 2013;471(6):2000–2006.

3. Weiss RJ, Ekström W, Hansen BH, et al. **Pathological subtrochanteric fractures in 194 patients: a comparison of outcome after surgical treatment of pathological and non-pathological fractures.** *J Surg Oncol.* 2013;107(5):498–504.

4. Cheng EY. **Prospective quality of life research in bony metastatic disease.** *Clin Orthop Relat Res.* 2003;(415 Suppl):S289-97.

5. Sneeuw KCA, Albertsen PC, Aaronson NK. **Comparison of patient and spouse assessments of health related quality of life in men with metastatic prostate cancer.** *J Urol* 2001;165(2):478–82.

6. Stephens RJ, Hopwood P, Girling DJ, et al. **Randomized trials with quality of life endpoints: are doctors' ratings of patients' physical symptoms interchangeable with patients' self-ratings?** *Qual Life Res.* 1997;6(3):225–36.

7. Wilson KA, Dowling AJ, Abdolell M, et al. **Perception of quality of life by patients, partners and treating physicians.** *Qual Life Res.* 2000;9(9):1041–1052.

8. Fasse L, Flahault C, Brédart A, et al. **Describing and understanding depression in spouses of cancer patients in palliative phase.** *Psychooncology.* 2015;24(9):1131–7.

9. Geng H, Chuang D, Yang F, et al. **Prevalence and determinants of depression in caregivers of cancer patients.** *Medicine (Baltimore).* 2018;97(39):e11863.

10. Katagiri H, Okada R, Takagi T, et al. **New prognostic factors and scoring system for patients with skeletal metastasis.** *Cancer Med.* 2014;3(5):1359–1367.

11. Quan H, Li B, Couris CM, et al. **Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries.** *Am J Epidemiol.* 2011;173(6):676–682.

12. Herdman M, Gudex C, Lloyd A, et al. **Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L).** *Qual Life Res.* 2011;20(10):1727–36.

13. Ader DN. **Developing the Patient-Reported Outcomes Measurement Information System (PROMIS).** *Med Care.* 2007;45(Suppl 1):S1–S2.

14. Likert R. **A technique for the measurement of attitudes.** *Arch Psychol.* 1932;140:1–55.

15. Liu H, Cella D, Gershon R, et al. **Representativeness of the atient-Reported Outcomes Measurement Information System internet panel.** *J Clin Epidemiol.* 2010;63(11):1169–78.

16. Yost KJ, Eton DT, Garcia SF, et al. **Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients.** *J Clin Epidemiol.* 2011;64(5):507–516.

17. Pickard AS, Neary MP, Cella D. **Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer.** *Health Qual Life Outcomes.* 2007;5(1):70.

18. Oken M, Creech R, Tormey D, et al. **Toxicity and response criteria of the Eastern Cooperative Oncology Group.** *Am J Clin Oncol.* 1982;5(6):649–656.

19. Jones JM, McPherson CJ, Zimmermann C, et al. Assessing agreement between terminally ill cancer patients' reports of their quality of life and family caregiver and palliative care physician proxy ratings. *J Pain Symptom Manage.* 2011;42(3):354–65.

20. Sneeuw KCA, Aaronson NK, Sprangers MAG, et al. Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. *J Clin Epidemiol.* 1998;51(7):617–631.

21. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)-a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009;42(2):377–81.

22. Cella DF, Tulsky DS. Quality of life in cancer: definition, purpose, and method of measurement. *Cancer Invest.* 1993;11(3):327–336.

23. Litwin MS, Lubeck DP, Henning JM, et al. Differences in urologist and patient assessments of health related quality of life in men with prostate cancer: results of the CaPSURE database. *J Urol.* 1998;159(6):1988–92.

24. Pearcy R, Waldron D, O'Boyle C, et al. Proxy assessment of quality of life in patients with prostate cancer: how accurate are partners and urologists? *J R Soc Med.* 2008;101(3):133–138.

25. Sneeuw KCA, Aronson NK, Sprangers MAG, et al. Evaluating the quality of life of cancer patients: assessments by patients, significant others, physicians and nurses. *Br J Cancer.* 1999;81(1):87–94.

26. Epstein AM, Hall JA, Tognetti J, et al. Using proxies to evaluate quality of life. Can they provide valid information about patients' health status and satisfaction with medical care? *Med Care.* 1989;27(3 Suppl):S91-8.

27. O'Brien J, Francis A. The use of next-of-kin to estimate pain in cancer patients. *Pain.* 1988;35(2):171–8.

28. Sprangers MA, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. *J Clin Epidemiol.* 1992;45(7):743–60.

29. Sneeuw KCA, Aaronson NK, Sprangers MA, et al. Value of caregiver ratings in evaluating the quality of life of patients with cancer. *J Clin Oncol.* 1997;15(3):1206–17.

30. Hopwood P, Stephens RJ, Machin D. Approaches to the analysis of quality of life data: experiences gained from a medical research council lung cancer working party palliative chemotherapy trial. *Qual Life Res.* 1994;3(5):339–52.

31. Ji J, Zöller B, Sundquist K, et al. Increased risks of coronary heart disease and stroke among spousal caregivers of cancer patients. *Circulation.* 2012;125(14):1742–7.

32. Miller GE, Murphy MLM, Cashman R, et al. Greater inflammatory activity and blunted glucocorticoid signaling in monocytes of chronically stressed caregivers. *Brain Behav Immun.* 2014;41:191–9.

33. Vitaliano PP, Zhang J, Scanlan JM. Is caregiving hazardous to one's physical health? A meta-analysis. *Psychol Bull.* 2003;129(6):946–72.

34. Blazeby JM, Williams MH, Alderson D, et al. Observer variation in assessment of quality of life in patients with oesophageal cancer. *Br J Surg.* 1995;82(9):1200–3.

# SUPPLEMENTAL MATERIAL TO CHAPTER 5

**Appendix 1.** Flowchart of enrolled patients.

**Appendix 2.** Baseline comparison between included and excluded patients

Supplemental material can be consulted online per the website of the journal and/or publisher.

PART III

# MORTALITY AND COMPLICATIONS

CHAPTER

06

# CLINICAL OUTCOME DIFFERENCES IN THE TREATMENT OF IMPENDING VERSUS COMPLETED PATHOLOGICAL LONG BONE FRACTURES

Olivier Q. Groot, Amanda Lans, Peter K. Twining, Michiel E.R. Bongers, Neal D. Kapoor, Jorrit-Jan Verlaan, Erik T. Newman, Kevin A. Raskin, Santiago A. Lozano-Calderon, Stein J. Janssen, Joseph H. Schwab

*The Journal of Bone & Joint Surgery 2021*

# ABSTRACT

### Background

There is a paucity of data about the benefits of prophylactic surgery in patients with long bone metastases. No randomized data exists for obvious ethical reasons, and thus support for the use of prophylactic surgery has largely been extrapolated based on relatively small sample sizes or are based on registry data which often insufficiently control for potential confounders.

### Objectives

We aimed to assess differences in outcome between surgical treatment of impending versus completed pathological fractures in long bone metastases for:

(1) 90-day and 1-year survival and

(2) intraoperative blood loss, perioperative blood transfusion, anesthesia time, duration of hospitalization, 30-day postoperative systemic complications, and reoperations.

### Design

Retrospective propensity score matched cohort study.

### Methods

We retrospectively performed a matched cohort study utilizing a database of 1,064 operative patients including 462 impending fractures and 602 completed metastatic long bone fractures. After matching on 22 variables, including primary tumor, visceral metastases, and surgical treatment, 270 impending pathological fractures were matched to 270 completed pathological fractures. The primary outcome –90-day and 1-year survival– was assessed by the cox proportional hazard model. The secondary outcomes were assessed by the McNemar test for categorical and Wilcoxon signed-rank test for continuous outcomes.

### Results

The 90-day survival did not differ between groups (HR 1.13, 95%CI 0.81-1.56, P=0.48), while there was an improved 1-year survival for impending fractures (38%vs.46%; HR 1.28, 95%CI 1.02-1.61, P=0.03). Regarding secondary outcomes, impending fractures had lower intraoperative estimated blood loss (P=0.03); lower rate of perioperative blood transfusions (P=0.01); shorter anesthesia time (P=0.04); and patients underwent fewer reoperations (OR 2.50, 95%CI 1.92-7.86, P=0.049), while we found no differences for 30-days postoperative complications or hospitalization duration.

**Conclusion**

Patients undergoing surgery for impending pathological fractures have lower 1-year mortality rates and better secondary outcomes as compared with patients undergoing surgery for completed pathological fractures, while accounting for 22 confounders through propensity matching. Patients with an impending pathological fracture are suggested to benefit from prophylactic stabilization as stabilizing a completed pathological fracture seems to be associated with increased mortality, blood loss, rate of blood transfusions, duration of surgery, and reoperation risk.

# INTRODUCTION

Skeletal metastases compromise the structural integrity of involved bone, leading to an increased pathological fracture risk.[1] Pathological fractures can result in significant morbidity and loss of quality of life.[2–4] When a metastatic lesion is at substantial risk of fracture, prophylactic stabilization is often advised to avert additional morbidity. Prophylactic surgery may be technically easier and allows for the consideration of multiple surgical options, some of which may not be feasible in the setting of a completed fracture. It also allows for preoperative work-up and optimization, as well as timing with respect to systemic therapy.[3] This avoids the potential "traumatic" morbidity of a completed fracture, including for example a fall and hematoma formation.

Prior studies suggest that prophylactic fixation of an impending pathological fracture is associated with lower levels of postoperative pain, lower complication rate, faster rehabilitation, and improved survival.[5–12] However, most studies are limited by relatively small sample sizes or are based on registry data which often insufficiently control for potential confounders. Propensity score matching is a statistical technique that limits the inherent shortcomings of non-experimental study designs by generating comparable distributions of relevant variables to reduce confounding.[13–15]

The purpose of this study was therefore to assess differences for: (1) 90-day and 1-year survival, and (2) intraoperative blood loss, perioperative blood transfusion, anesthesia time, duration of hospitalization, 30-day postoperative systemic complications, and reoperations between surgically treated impending pathological fractures and completed pathological fractures in patients with long bone metastases.

# METHODS

### Patient Cohort

Our institutional review board approved a waiver of informed consent for this retrospective propensity score matched cohort study. This study was performed at two urban tertiary care referral centers for orthopaedic oncology in the United States affiliated within one healthcare entity.

We included all 1,064 consecutive adult (18 years or older) patients that underwent surgery between 1999 and 2017 for an impending pathological or completed pathological fracture due to a long bone metastasis (Figure 1).[16] We defined long bones as: femur, humerus, tibia, fibula, radius, and ulna. Metastatic disease included skeletal metastases from solid tumors and sites of bony involvement in cases of multiple myeloma and lymphoma.[17] Exclusion criteria were: (1) revision procedures; (2) metastases from sarcoma; (3) pathological fractures in multiple bones requiring simultaneous surgery; and (4) surgery other than intramedullary nailing, dynamic hip screw fixation, plate-screw fixation, endoprosthetic reconstruction, or a combination thereof. Sarcoma was excluded as we considered sarcoma metastases treatment to be substantially different. Additionally, the large number of sarcomas treated at the included tertiary centers would limit generalizability of our findings. If a patient had more than one qualifying surgery during the study period, only the first surgery was included to avoid violating the statistical law of independence. Choice of treatment was decided by mutual agreement between the patient and surgeon. In general, the Mirels score was used to estimate fracture risk, and prophylactic fixation was recommended in patients with a score of eight or higher.[18] During the study period, postoperative care and rehabilitation was tailored to disease severity.

### Outcomes and Explanatory Variables

The primary outcome measures were 90-day and 1-year survival after surgery. Date of death was determined using the Social Security Death Index and by reviewing medical records. Loss to follow-up was 3% (33 of 1,064) at 90-days and 6% (60 of 1,064) at 1-year. The secondary outcome measures were: (1) intraoperative blood loss (liters); (2) perioperative allogeneic blood transfusion (transfusion of packed red blood cells 7 days prior to and up to 7 days after surgery); (3) anesthesia time (hours); (4) duration of hospitalization (days); (5) systemic postoperative complications within 30 days; and (6) local reoperation to the surgical site (only the first reoperation was accounted for). We considered the following postoperative complications within 30 days: pneumonia, venous thromboembolism, sepsis, myocardial infarction, wound infection and/or dehiscence, and urinary tract infection.[19–21]

Factors known or suggested to be associated with survival were included as explanatory variables.[5–8,10–12,17,22,23] Medical records were manually reviewed to obtain these variables: age; sex; BMI

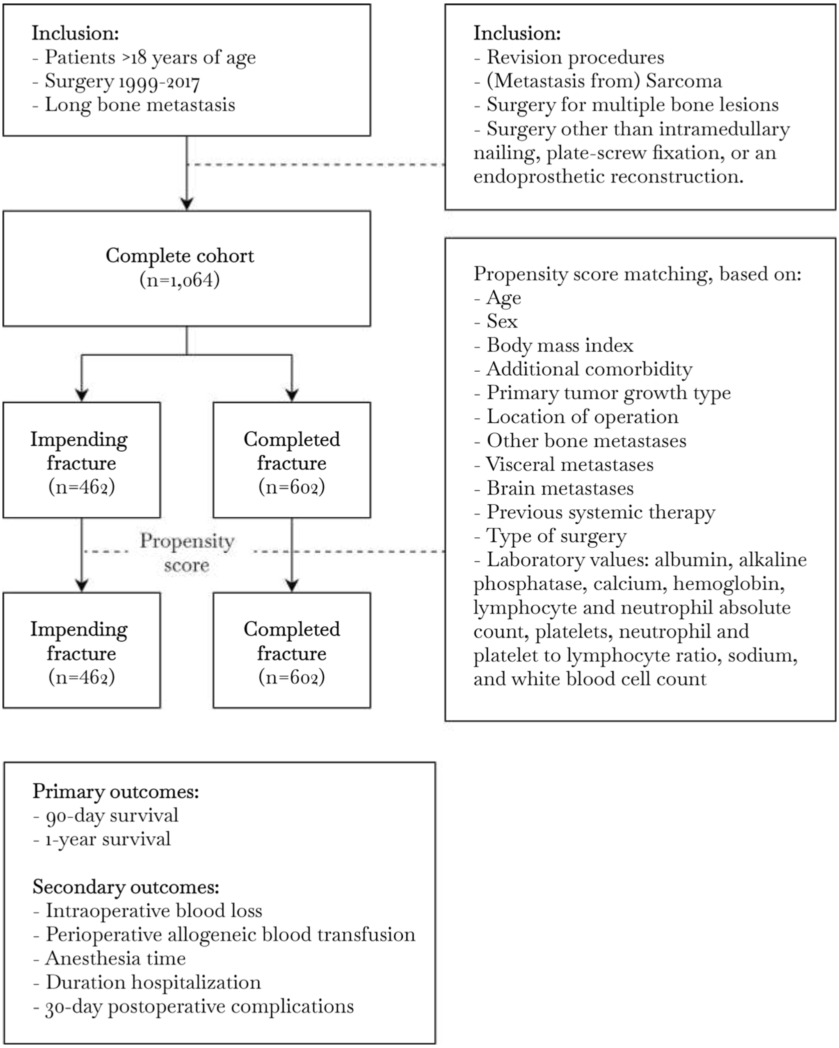

Inclusion:
- Patients >18 years of age
- Surgery 1999-2017
- Long bone metastasis
Inclusion:
- Revision procedures
- (Metastasis from) Sarcoma
- Surgery for multiple bone lesions
- Surgery other than intramedullary nailing, plate-screw fixation, or an endoprosthetic reconstruction.
Complete cohort
(n=1,064)
Propensity score matching, based on:
- Age
- Sex
- Body mass index
- Additional comorbidity
- Primary tumor growth type
- Location of operation
- Other bone metastases
- Visceral metastases
- Brain metastases
- Previous systemic therapy
- Type of surgery
- Laboratory values: albumin, alkaline phosphatase, calcium, hemoglobin, lymphocyte and neutrophil absolute count, platelets, neutrophil and platelet to lymphocyte ratio, sodium, and white blood cell count
Impending fracture
(n=462)
Completed fracture
(n=602)
Propensity score
Impending fracture
(n=462)
Completed fracture
(n=602)
Primary outcomes:
- 90-day survival
- 1-year survival
Secondary outcomes:
- Intraoperative blood loss
- Perioperative allogeneic blood transfusion
- Anesthesia time
- Duration hospitalization
- 30-day postoperative complications

**Figure 1.** Flow diagram illustrating patient selection and matching process.

([body mass index], kg/m$^2$); any Charlson comorbidity in addition to metastatic cancer; primary tumor type categorized as slow, moderate or rapid growth as classified by Katagiri et al.; tumor location; additional bone metastases; visceral metastases (lung and/or liver); brain metastases; previous systemic therapy; type of surgical treatment, and eight preoperative laboratory values, nearest to surgery with a maximum of 7 days.[24,25] Missing data are displayed in Table 1 and were imputed using single median imputation prior to propensity score matching. Bivariate analyses were completed case analyses.

### Statistical Analysis

Nonparametric testing was used for continuous variables as some variables had skewed distributions. In bivariate analysis before matching, baseline characteristics were compared between patients with impending and completed fracture using the Mann-Whitney U test for continuous variables and Fisher Exact test for categorical variables.

Propensity score matching was used to generate comparable cohorts with a similar distribution of covariates by matching on variables known to be associated with survival in patients with long bone metastases.[13] Propensity score matching was conducted using a one-to-one nearest-neighbor matching in a random order without replacement and with a caliper fixed at 0.005 (maximum allowable difference in propensity scores) based on a propensity score calculated through a logit model including all explanatory variables. Only patients matched with propensity scores were included in the analyses. Using this technique, 270 impending fracture cases were matched to 270 completed fracture cases. The adequacy of matching was then demonstrated by: (1) testing the standardized mean differences (SMD); (2) comparing the matched variables using the Wilcoxon signed-rank test and McNemar test for continuous and categorical variables, respectively; and (3) a Kernel density plot.14 After propensity score matching, the matched groups did not differ for any of the explanatory variables (P>0.05), and none of the differences were substantial (>0.25) as demonstrated by standardized mean differences (Table 1). Kernel density plots demonstrated adequate matching (Figure 2).

The primary outcome – 90-day and 1-year survival – was tested between the matched groups using six different methods to consolidate the strength of our findings. First, the log-rank test compared

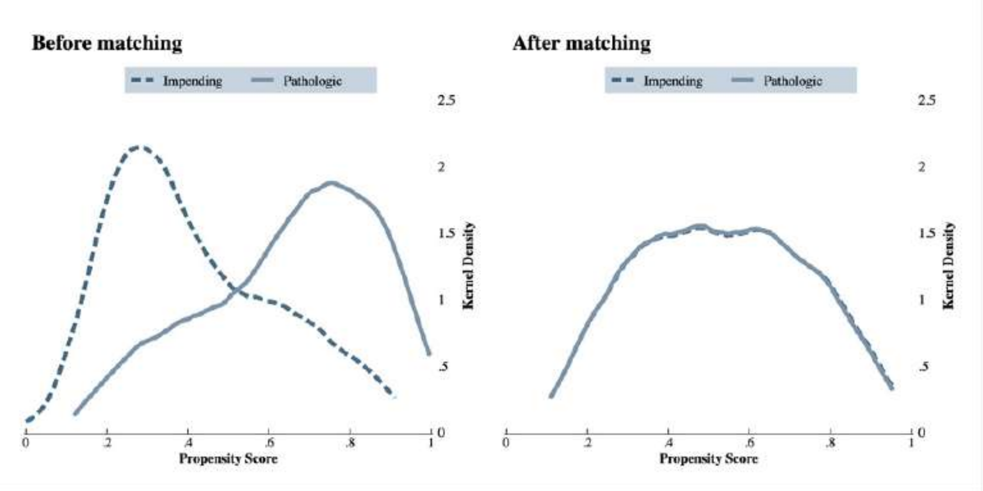


**Figure 2.** Kernel density plots demonstrate the distribution of the propensity score before and after matching, demonstrating the adequateness of propensity score matching.

the equality of survival curves, stratified by propensity score matched pairs. Second, the McNemar test compared the matched pairs on a dichotomous predictor (impending versus completed fracture) and dichotomous outcome (deceased or not). Third, four different Cox proportional hazard models were used: (1) unadjusted; (2) stratified into five quintiles by their propensity scores and the average of each quintile stratum was taken; (3) robust variance estimator; and (4) weighted by the inverse probability of treatment (IPT) using the propensity score.[14,15] Kaplan-Meier plots demonstrate the survival curves for both groups before and after propensity score matching.

The secondary outcomes were assessed using paired tests; McNemar for dichotomous outcomes, and Wilcoxon signed rank for continuous data. Odd ratios (OR) and hazard ratios (HR) with 95% confidence intervals (CI) were calculated. A two-tailed P-value of < 0.05 was considered significant. All statistical analyses were performed using Stata 15.0 (StataCorp LP, College Station, TX, USA).

# RESULTS

### 90-Day and 1-Year Survival

After propensity score matching, the 90-day survival did not differ between the impending and completed fracture groups with a survival rate of 73% (197/270) in the impending fracture group and 71% (193/270) in the completed fracture group (HR 1.13, 95%CI 0.81-1.56, *P*=0.48). The 1-year survival rate was higher in the impending fracture group with 46% (126/270) compared with 38% (102/270) in the completed fracture group (Cox proportional hazard model weighted by IPT; HR 1.28, 95%CI 1.02-1.61, *P*=0.03; Figure 3). Unadjusted, stratified by quintiles, and robust variance estimator Cox hazard models yielded comparable results (Appendix 1).

### Secondary Outcomes

After propensity score matching, the impending fracture group had lower intraoperative blood loss in the impending fracture group with a median of 0.2 liters (IQR 0.1-0.4) compared with a median of 0.3 liters (IQR 0.2-0.4) in the completed fracture group (*P*=0.03); less blood transfusions with a median of 0 transfusions (IQR 0-2) compared with a median of 1 transfusion (IQR 0-2) in the completed fracture group (*P*=0.01); shorter anesthesia time in the impending fracture group with a median of 2.8 hours (IQR 2.1-3.5) compared with a median of 3.1 hours (IQR 2.5-3.6) in the completed fracture group (*P*=0.04); and fewer reoperations with 3.3% (9/270) compared with 6.7% (18/270) in the completed fracture group (OR 2.50, 95%CI 1.92-7.86, *P*=0.049); The duration of hospitalization and rate of systemic postoperative complications within 30 days did not differ between the impending and completed fracture groups; median duration of hospitalization was 4 days (IQR 3-7) in both groups (*P*=0.09); 30-day systemic complication rate was 14% (38/270) in the impending fracture group and 16% (42/270) in the completed fracture group (OR 1.12, 95%CI 0.69-1.83, *P*=0.64).

**Table 1.** Comparison of baseline characteristics between impending and completed

| | Before propensity score matching (n=1,064) | | | |
|---|---|---|---|---|
| | Impending (n=462) | Completed (n=602) | | |
| | *Median (IQR)* | *Median (IQR)* | *P-value* | *Std. Diff.* |
| Age (years) | 61 (53-70) | 64 (56-72) | **<0.01** | -0.17 |
| Body mass index (in kg/$m^2$)[a] | 27 (23-30) | 27 (23-30) | 0.94 | -0.01 |
| Preoperative Laboratory values[a] | | | | |
| Albumin (g/dL) | 3.8 (3.3-4.2) | 3.6 (3.2-4.0) | **<0.01** | 0.26 |
| Alkaline phosphatase (IU/L) | 99 (73-131) | 105 (75-156) | 0.06 | -0.19 |
| Calcium (mg/dL) | 9.2 (8.8-9.7) | 9.1 (8.7-9.6) | **0.01** | 0.17 |
| Hemoglobin (g/dL) | 12 (10-13) | 11 (10-12) | **<0.01** | 0.24 |
| Lymphocyte absolute count ($10^3$/ L) | 1.1 (0.7-1.6) | 1.0 (0.6-1.5) | **0.02** | 0.11 |
| Neutrophil absolute count ($10^3$/ L) | 5.0 (3.5-7.3) | 5.8 (3.9-8.2) | **<0.01** | -0.25 |
| Neutrophil to lymphocyte ratio | 4.7 (3.0-7.4) | 5.7 (3.2-9.8) | **<0.01** | -0.28 |
| Platelet count ($10^3$/$mm^3$) | 259 (199-343) | 241 (174-322) | **<0.01** | 0.20 |
| Platelet to lymphocyte ratio | 230 (158-370) | 239 (160-383) | 0.31 | -0.10 |
| Sodium (mg/dL) | 138 (136-140) | 138 (135-140) | **0.01** | 0.20 |
| White blood cell count ($10^3$/ L) | 7.2 (5.1-9.5) | 7.5 (5.2-10) | 0.11 | -0.11 |
| | n (%) | n (%) | | |
| Female | 262 (57) | 333 (55) | 0.66 | -0.03 |
| Additional comorbidity[b] | 245 (53) | 331 (55) | 0.54 | -0.04 |
| Primary Tumor Growth[c] | | | **0.01** | 0.18 |
| Slow | 174 (38) | 280 (47) | | |
| Moderate | 112 (24) | 134 (22) | | |
| Rapid | 176 (38) | 188 (31) | | |
| Tumor location | | | **<0.01** | 0.57 |
| Upper extremity | 49 (11) | 201 (33) | | |
| Lower extremity | 413 (89) | 401 (67) | | |
| Other bone metastases[d] | 355 (77) | 466 (77) | 0.83 | -0.01 |
| Visceral metastases | 217 (47) | 258 (43) | 0.19 | 0.08 |
| Brain metastases | 89 (19) | 82 (14) | **0.02** | 0.15 |
| Previous systemic therapy | 289 (63) | 372 (62) | 0.85 | 0.02 |
| Type of surgery | | | **<0.01** | -0.38 |
| Intramedullary nail | 355 (77) | 269 (45) | | |
| Endoprosthetic reconstruction | 37 (8.0) | 203 (34) | | |
| Plate and screw fixation | 46 (10) | 107 (18) | | |
| Dynamic hip screw | 10 (2.2) | 9 (1.5) | | |
| Multiple implants | 14 (3.0) | 14 (2.3) | | |

*IQR=interquartile range; Std. Diff.=standardized difference; mL=milliliter; g/dL=gram per deciliter;  L=microliter; mg/dL=milligram per deciliter; $mm^3$=cubic millimeter; kg/$m^2$=kilogram per square meter.* **Bold** *indicates significance ($P<0.05$).*

*a Patient data was available for respectively impending and completed pathological fracture: BMI 375 (81%) and 458 (76%), albumin 313 (68%) and 441 (73%), alkaline phosphatase 317 (69%) and 439 (73%), calcium 370 (80%) and 498 (83%), hemoglobin 392 (85%) and 529 (88%), lymphocyte absolute count 318 (69%) and 428 (71%), neutrophil absolute count 322 (70%) and 428 (71%), neutrophil to lymphocyte ratio 318 (69%) and 428 (71%), platelet count 393 (85%) and 528 (88%), platelet to lymphocyte ratio 318 (69%) and 427 (71%), sodium 365 (79%) and 504 (84%), and white blood cell count 392 (85%) and 529 (88%).*

*b These values were based on any additional comorbidity on top of the metastatic disease score according to the modified Charlson Comorbidity Index.*

pathological fracture before (n=1,064) and after (n=540) propensity score matching.

| After propensity score matching (n=540) | | | |
|---|---|---|---|
| Impending (n=270) | Completed (n=270) | | |
| *Median (IQR)* | *Median (IQR)* | *P-value* | *Std. Diff.* |
| 63 (54-71) | 63 (53-71) | 0.95 | 0.03 |
| 27 (24-29) | 27 (24-29) | 0.81 | 0.02 |
| | | | |
| 4.1 (3.6-4.7) | 4.0 (3.5-4.7) | 0.75 | 0.04 |
| 101 (80-121) | 101 (87-120) | 0.30 | -0.04 |
| 9.2 (8.9-9.6) | 9.2 (8.9-9.6) | 0.99 | 0.00 |
| 11 (10-13) | 11 (10-12) | 0.39 | 0.02 |
| 1.0 (0.8-1.3) | 1.0 (0.8-1.2) | 0.35 | 0.04 |
| 5.5 (4.1-6.6) | 5.5 (4.5-6.9) | 0.39 | -0.04 |
| 5.4 (3.8-6.2) | 5.4 (4.2-6.7) | 0.60 | -0.04 |
| 251 (204-308) | 251 (199-332) | 0.46 | -0.07 |
| 250 (179-320) | 250 (186-344) | 0.38 | -0.05 |
| 138 (137-139) | 138 (136-139) | 0.26 | 0.08 |
| 7.3 (5.6-9.4) | 7.3 (5.6-9.7) | 0.76 | -0.02 |
| n (%) | n (%) | | |
| 158 (59) | 161 (60) | 0.79 | 0.02 |
| 149 (55) | 144 (53) | 0.67 | 0.04 |
| | | 0.24 | -0.11 |
| 118 (44) | 107 (40) | | |
| 64 (24) | 60 (22) | | |
| 88 (33) | 103 (38) | | |
| | | 0.99 | 0.00 |
| 47 (17) | 47 (17) | | |
| 223 (83) | 223 (83) | | |
| 212 (79) | 216 (80) | 0.68 | -0.04 |
| 120 (44) | 134 (50) | 0.25 | -0.10 |
| 48 (18) | 48 (18) | 1.00 | 0.00 |
| 175 (65) | 179 (66) | 0.72 | -0.03 |
| | | 0.99 | 0.24 |
| 168 (62) | 169 (63) | | |
| 36 (13) | 73 (27) | | |
| 45 (17) | 23 (8.5) | | |
| 9 (3.3) | 1 (0.3) | | |
| 12 (4.4) | 4 (1.5) | | |

*c Based on histology groupings; slow growth includes hormone dependent breast cancer, hormone dependent prostate cancer malignant lymphoma malignant myeloma, and thyroid cancer; moderate growth includes non-small cell lung cancer with molecularly targeted therapy, hormone independent breast cancer, hormone independent prostate cancer, renal cell carcinoma, sarcoma, other gynecological cancer, and others; and rapid growth includes other lung cancer, colon and rectal cancer, gastric cancer, hepatocellular carcinoma, pancreatic cancer, head and neck cancer, other urological cancer, esophageal cancer, malignant melanoma, gallbladder cancer, cervical cancer, and unknown origin. When testing primary tumor type distribution after propensity score matching, we found no difference between groups (p=0.59).*
*d Any bone metastasis outside of the lesion treated for.*

Table 2. Comparison of primary and secondary outcomes in patients with impending and matching.

| | Before propensity score matching (n=1,064) | | | | |
|---|---|---|---|---|---|
| | Impending (n=462) | Completed (n=602) | | Standard | |
| | | *n (%)* | *HR (95% CI)* | *error* | *P-value* |
| Survival[a] | | | | | |
| 90-days | 341 (74) | 424 (70) | 1.17 (0.93-1.48) | 0.139 | 0.17 |
| 1-year | 202 (44) | 236 (39) | 1.16 (0.99-1.36) | 0.094 | 0.07 |
| | *Median (IQR)/n (%)* | | *OR (95% CI)* | | |
| Intraoperative blood loss (liters)[a] | 0.2 (0.1-0.3) | 0.3 (0.2-0.5) | - | - | **<0.01** |
| Perioperative allogeneic blood transfusion | 0 (0-2) | 1 (0-3) | - | - | **<0.01** |
| Anesthesia time (hours)[a] | 2.8 (2.2-3.5) | 3.1 (2.5-3.8) | - | - | **<0.01** |
| Duration hospitalization (days)[a] | 4 (3-6) | 5 (3-7) | - | - | **<0.01** |
| Systemic postoperative complications within 30 days | 66 (14) | 83 (14) | 0.96 (0.68-1.36) | 0.171 | 0.82 |
| Reoperations | 16 (3.5) | 44 (7.3) | 2.20 (1.22-3.95) | 0.657 | **0.01** |

*IQR=interquartile range; Std. Diff.=standardized difference; CI=confidence interval; HR=hazard ratio; OR=odds ratio.* Bold *indicates significance (P<0.05).*

*a Patient data before propensity score matching was available for respectively impending and completed pathological fracture: survival 90-days 447 (97%) and 584 (97%), 1-year 436 (94%) and 568 (94%), intraoperative blood loss 408 (88%) and 517 (86%), anesthesia time 365 (79%) and 493 (82%), and hospitalization 456 (99%) and 585 (97%).*

*Patient data after propensity score matching was available for respectively impending and completed pathologic fracture: survival 90-days 262 (97%) and 262 (97%), 1-year 256 (95%) and 253 (94%), intraoperative blood loss 233 (86%) and 238 (88%), anesthesia time 210 (78%) and 222 (82%), and hospitalization 267 (99%) and 261 (97%). Both outcomes in matched pairs were available in estimated blood loss 203 (75%), anesthesia time 175 (65%), and hospitalization 258 (96%).*

# DISCUSSION

Metastatic bone disease can lead to pain, disability, and risk of development of a pathological fracture, which is associated with further deterioration in quality of life and possibly worse prognosis. Several studies have suggested improved outcome after prophylactic fixation of an impending fracture as compared with an acute pathological fracture; however, these studies were limited by small sample size or based on registry data with insufficient controlling for confounding.[5–12] Our relatively large study, using propensity score matching to create comparable cohorts across 22 explanatory variables, found that patients who underwent surgery for an impending pathological fracture had better 1-year survival, less intraoperative blood loss, fewer perioperative blood transfusions, shorter anesthesia time, and fewer reoperations in comparison with patients who underwent surgery for a completed pathological fracture. No differences were found for 90-day survival, 30-day systemic postoperative complications, and length of hospitalization between the two groups.

This study has several limitations. First, this was a retrospective study from medical centers affiliated within one healthcare entity, causing the inevitable risk of selection and confounding bias. To correct this, propensity-matching analysis was used. An experimental study design –such as a randomized controlled trial– is not completed pathological fractures before (n=1,064) and after (n=540) propensity score

| After propensity score matching (n=540) | | | | |
|---|---|---|---|---|
| Impending<br>(n=270) | Completed<br>(n=270) | | | |
| *n (%)* | | *HR (95% CI)*[b] | *Standard error* | *P-value* |
| | | | | |
| 197 (73) | 193 (71) | 1.13 (0.81-1.56) | 0.188 | 0.48 |
| 123 (46) | 102 (38) | 1.28 (1.02-1.61) | 0.148 | **0.03** |
| *Median (IQR)/n (%)* | | *OR (95% CI)* | | |
| 0.2 (0.1-0.4) | 0.3 (0.2-0.4) | - | - | **0.03** |
| 0 (0-2) | 1 (0-2) | - | - | **0.01** |
| 2.8 (2.1-3.5) | 3.1 (2.5-3.6) | - | - | **0.04** |
| 4 (3-7) | 4 (3-7) | - | - | 0.09 |
| 38 (14) | 42 (16) | 1.12 (0.69-1.83) | - | 0.64 |
| 9 (3.3) | 18 (6.7) | 2.50 (1.92-7.86) | - | **0.05** |

*[b] The presented hazard ratio after matching are based on the cox proportional hazard model weighted by the inverse probability of treatment weighting (IPT) using the propensity score. Additional survival analyses can be found in Appendix 1.*

possible and considered unethical for the study topic. Second, quality of life outcomes were not recorded during standard case visits which would have been a valuable addition to this study with a frail patient population. Third, estimated blood loss was based on anesthesia reports, whereas measuring hemoglobin balance is a more accurate method; however, this was not consistently measured in our cohort. Despite this potential inaccuracy, we do not feel that this was affected based on the fracture group and therefore believe that the significant difference is valid. Fourth, we were unable to account for patients who abstained from prophylactic surgery due to compelling factors as this was not documented uniformly. Fifth, propensity matching on specific systemic therapy data and post-surgery strategies is limited by diverse regimens and their change over time. Sixth, a recent study found an association between CRP and lower 1-year survival, indicating a potential important confounder to account for.[26] Unfortunately, we were unable to include this covariate in our propensity score matching model due to insufficient data. Lastly, we were unable to account for performance status in propensity score matching (e.g., ECOG/Karnofsky score), as these scores were available in less than half the patients. When analyzing available data; we found no significant difference in dichotomized ECOG score between groups after propensity score matching (fracture: ECOG 3-4=17% [n=16], ECOG 0-2=83% [n=91], missing=60% [n=161]; impending: ECOG 3-4=12% [n=13], ECOG

0-2=88% [n=92], missing=61% [n=165]; *P*=0.65).

The 90-day survival did not differ between the impending and completed pathological fracture groups in our study. Two previous studies –both of which used the same US based registry (NSQIP) during the same time period– also found no difference in 30-day survival (OR 2.38, 95%CI 0.88-6.25, *P*=0.09 among 1,317 patients with long bone metastases; and OR 1.71, 95%CI 0.95-3.09, *P*=0.07, among 620 femoral metastases only; Appendix 2 and 3).[7,8] Our long-term (1-year) survival was 8% better in patients that underwent prophylactic stabilization compared with acute stabilization of a pathological fracture. Ward et al. found a similar difference in 1-year survival rate in a single institutional cohort of 182 patients, but did not control for confounding factors (35% 1-year survival in impending fractures versus 25% in completed fractures, *P*=0.02).[12] In addition, three other studies based on registry data investigated overall survival in femoral lesions only, all demonstrated improved long-term survival in the impending fracture group as compared with the completed fracture group.[6,10,11] Overall, long-term survival is generally poor across patients surgically treated for metastatic bone disease regardless of whether the surgery was prophylactic or for an acute pathologic fracture. However, our results –supported by the named previous studies– suggest no difference in survival in the short-term but does demonstrate that patients have worse long-term survival when they developed a fracture. This might be related to the perioperative timeframe that may be pivotal in determining long-term cancer outcomes or the functional disabilities and the period of immobilization following a completed fracture.[1,27]

Our finding that prophylactic fixation was associated with lower rates of perioperative blood loss and less blood transfusions is in line with all three previous studies related to this topic. Both McLynn et al. and Aneja et al. found in registry data (NSQIP and NIS) of femoral metastases a similar decreased risk of blood transfusion in impending fractures (OR 0.62, 95%CI 0.38-0.89, *P*=0.01 among 620 patients; and OR 0.74, 95%CI 0.65-0.84, *P*<0.01 among 5,579 patients).[5,7] Only Ward et al. investigated intraoperative blood loss and found that there was less average blood loss in patients treated with prophylactic surgery in comparison to patients who sustained a completed fracture (438cc versus 636cc among 182 patients, *P*=0.01).[12] Increased transfusions have been reported to have an immunosuppressive effect which in turn might lead to worse survival.[19] This immunosuppressive effect offers another possible explanation as to why patients who were treated for a completed fracture had a decreased 1-year survival compared with patients treated for an impending fracture.[27]

Anesthesia time was shorter in patients treated for impending fractures. Arvinius et al, reported similar results, although they included only 65 patients and did not account for confounding factors (impending fracture: 23 minutes vs completed fracture 48 minutes; *P*=0.003).[10] However, McLynn et al. did not find a difference in surgery time in 620 femoral metastases using registry data (OR 1.31, 95%CI 0.90-1.90, *P*=0.16).[7]

No difference was found in duration of hospitalization. Multiple studies described hospitalization, with the majority suggesting a shorter hospital stay in the impending facture group.[7,8,10–12] For example, El Abiad et al. found in registry data of 1,317 patients that the impending fracture group had a shorter hospital stay compared with the completed fracture group (mean (standard deviation) of 6.9 (8.1) days versus 8.2 (9.0) days; $P$=0.01).[8] Earlier mobilization after prophylactic stabilization and a greater likelihood of being discharged to home may explain the shorter hospital stay in the prophylactic fracture group.

Systemic postoperative complications within 30 days did not differ between both groups. Prior studies have revealed mixed findings, although the majority suggest a higher complication rate in the prophylactic surgery group.[5,7,8,10,11] For example, El Abiad et al. found that prophylactic fixation was associated with a lower risk of major medical complications within 30 days after controlling for age, BMI, and disseminated cancer (OR 0.64, 95%CI, 0.45-0.93, $P$=0.02).[8] However, the studies that suggest a difference in complication rates use mostly registry based database and are subject to coding bias because complications are known to be miscoded by the physicians.[28]

Last, we found that less reoperations were performed in patients who were treated for an impending fracture. Only El Abiad et al. reported on reoperation rates within 30 days after surgery. Although using registry data and controlling for age, BMI, and disseminated cancer only, their results trended towards a similar difference (OR 0.65, 95%CI 0.42-1.01, $P$=0.06).[8] This may suggest that prophylactic surgical constructions are more stable and less prone to fail due to relative healthier local bone compared with a completed fracture. In addition, operating on a completed fracture is considered a more complex surgery due to fracture reduction and reconstruction and the possibility of more soft-tissue damage, which contribute to impaired surgical constructions compared with prophylactic surgery.

The correct and timely identification of metastatic bone lesions that is at risk for developing a completed pathological fracture, and significant morbidity to patients, is essential for physicians providing oncological care including radiation oncologists, orthopaedic oncologists, and medical oncologists. Accurate identification creates opportunity for prophylactic surgical stabilization, which seem to result in improved clinical outcomes. In addition, the limited survival of patients with metastatic bone disease must be considered when considering surgical stabilization to allow physicians and patients to make informed treatment decisions in line with their goals and expectations. Therefore, it is fundamental to correctly identify which lesions are causing disability and are at risk for developing a fracture to prevent patients from undergoing unnecessary surgical intervention. Currently available predictive models for fractures are limited by their inaccuracies and difficulty in use. For example, the widely known Mirels score has been shown to lack sufficient sensitivity and specificity, and interobserver agreement is moderate.[18,29] CT-based predictive algorithms developed by Snyder et al. show promising results, but clinical application might be limited due to selection bias and difficulty in use.[30,31] In order to benefit clinical oncologic practice, future research should aim to develop an accessible, easy to use and accurate prediction tool which identifies if a patient is at risk to develop a completed fracture. With this tool, patients who may benefit from prophylactic surgical stabilization can be identified.

## CONCLUSION

This retrospective propensity score matched study found that patients treated for an impending pathological fracture had better 1-year survival, less intraoperative blood loss, fewer perioperative blood transfusions, shorter anesthesia time, and fewer reoperations than patients treated for an completed pathological fracture of a metastatic long bone lesion. Choosing the optimal candidate for prophylactic surgery remains paramount to avoid overtreatment. For the advancement of clinical oncologic care, it will be helpful to develop an easy to use, accurate, and validated prediction tool which identifies if a patient with a metastatic bone lesion is at risk for developing a completed pathological fracture.

# REFERENCES

1. Coleman RE. Metastatic bone disease: clinical features, pathophysiology and treatment strategies. *Cancer Treat Rev.* 2001;27(3):165–176.

2. Damron TA, Mann KA. Fracture risk assessment and clinical decision making for patients with metastatic bone disease. *J Orthop Res Off Publ Orthop Res Soc.* 2020;38(6):1175–1190.

3. Coleman RE. Clinical features of metastatic bone disease and risk of skeletal morbidity. *Clin cancer Res an Off J Am Assoc Cancer Res.* 2006;12(20 Pt 2):6243s-6249s.

4. von Moos R, Body JJ, Egerdie B, et al. Pain and analgesic use associated with skeletal-related events in patients with advanced cancer and bone metastases. *Support care cancer Off J Multinatl Assoc Support Care Cancer.* 2016;24(3):1327–1337.

5. Aneja A, Jiang JJ, Cohen-Rosenblum A, et al. Thromboembolic disease in patients with metastatic femoral lesions: a comparison between prophylactic fixation and fracture fixation. *J Bone Joint Surg Am.* 2017;99(4):315–323.

6. Philipp TC, Mikula JD, Doung YC, et al. Is there an association between prophylactic femur stabilization and survival in patients with metastatic bone disease? *Clin Orthop Relat Res.* 2020;478(3):540–546.

7. McLynn RP, Ondeck NT, Grauer JN, et al. What is the adverse event profile after prophylactic treatment of femoral shaft or distal femur metastases? *Clin Orthop Relat Res.* 2018;476(12):2381–2388.

8. El Abiad JM, Raad M, Puvanesarajah V, et al. Prophylactic versus postfracture stabilization for metastatic lesions of the long bones: a comparison of 30-day postoperative outcomes. *J Am Acad Orthop Surg.* 2019;27(15):e709–e716.

9. Blank AT, Lerman DM, Patel NM, et al. Is prophylactic intervention more cost-effective than the treatment of pathologic fractures in metastatic bone disease? *Clin Orthop Relat Res.* 2016;474(7):1563–1570.

10. Arvinius C, Parra JLC, Mateo LS, et al. Benefits of early intramedullary nailing in femoral metastases. *Int Orthop.* 2014;38(1):129–132.

11. Ristevski B, Jenkinson RJ, Stephen DJG, et al. Mortality and complications following stabilization of femoral metastatic lesions: a population-based study of regional variation and outcome. *Can J Surg.* 2009;52(4):302–308.

12. Ward WG, Holsenbeck S, Dorey FJ, et al. Metastatic disease of the femur: surgical treatment. *Clin Orthop Relat Res.* 2003;(415 Suppl):S230-44.

13. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med.* 2009;28(25):3083–3107.

14. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med.* 2014;33(7):1242–1258.

15. Austin PC, Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study. *Stat Methods Med Res.* 2016;25(5):2214–2237.

16. Janssen SJ, van der Heijden AS, van Dijke M, et al. 2015 Marshall Urist Young investigator award: prognostication in patients with long bone metastases: does a boosting algorithm improve survival estimates? *Clin Orthop Relat Res.* 2015;473(10):3112–3121.

17. Nathan SS, Healey JH, Mellano D, et al. Survival in patients operated on for pathologic fracture: Implications for end-of-life orthopedic care. *J Clin Oncol.* 2005;23(25):6072–6082.

18. Mirels H. Metastatic disease in long bones. A proposed scoring system for diagnosing impending pathologic

fractures. *Clin Orthop Relat Res.* 1989;(249):256–264.

19. Janssen SJ, Braun Y, Ready JE, et al. Are allogeneic blood transfusions associated with decreased survival after surgery for long-bone metastatic fractures? *Clin Orthop Relat Res.* 2015;473(7):2343–2351.

20. Janssen SJ, Kortlever JTP, Ready JE, et al. Complications after surgical management of proximal femoral metastasis: A retrospective study of 417 patients. *J Am Acad Orthop. Surg.* 2016;24(7):483–494.

21. Groot OQ, Ogink PT, Janssen SJ, et al. High risk of venous thromboembolism after surgery for long bone metastases: A retrospective study of 682 patients. *Clin Orthop Relat Res.* 2018;476(10).

22. Willeumier JJ, van der Linden YM, van der Wal CWPG, et al. An easy-to-use prognostic model for survival estimation for patients with symptomatic long bone metastases. *J Bone Joint Surg Am.* 2018;100(3):196–204.

23. Forsberg JA, Eberhardt J, Boland PJ, et al. Estimating survival in patients with operable skeletal metastases: an application of a bayesian belief network. *PLoS One.* 2011;6(5):e19956.

24. Katagiri H, Okada R, Takagi T, et al. New prognostic factors and scoring system for patients with skeletal metastasis. *Cancer Med.* 2014;3(5):1359–1367.

25. Quan H, Li B, Couris CM, et al. Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol.* 2011;173(6):676–682.

26. Errani C, Cosentino M, Ciani G, et al. C-reactive protein and tumour diagnosis predict survival in patients treated surgically for long bone metastases. *Int Orthop.* 2021;45(5):1337–1346.

27. Horowitz M, Neeman E, Sharon E, et al. Exploiting the critical perioperative period to improve long-term cancer outcomes. *Nat Rev Clin Oncol.* 2015;12(4):213–226.

28. Yoshihara H, Yoneoka D. Understanding the statistics and limitations of large database analyses. *Spine (Phila. Pa. 1976).* 2014;39(16):1311–1312.

29. Damron TA, Morgan H, Prakash D, et al. Critical evaluation of Mirels' rating system for impending pathologic fractures. *Clin Orthop Relat Res.* 2003;(415 Suppl):S201-7.

30. Janssen SJ, Paulino Pereira NR, Meijs TA, et al. Predicting pathological fracture in femoral metastases using a clinical CT scan based algorithm: A case-control study. *J Orthop Sci.* 2018;23(2):394–402.

31. Snyder BD, Hauser-Kara DA, Hipp JA, et al. Predicting fracture through benign skeletal lesions with quantitative computed tomography. *J Bone Joint Surg Am.* 2006;88(1):55–70.

# SUPPLEMENTAL MATERIAL TO CHAPTER 6

**Appendix 1.** Comparison of different methods for survival analysis after propensity score matching (n=540).

**Appendix 2.** Summary of study characteristics that compare patients undergoing surgery for impending and completed pathological fracture in long bone metastases as their primary study aim.

**Appendix 3.** Summary of study outcomes that compare patients undergoing surgery for impending and completed pathological fracture in long bone metastases as their primary study aim.

Supplemental material can be consulted online per the website of the journal and/or publisher.

CHAPTER

07

# HIGH RISK OF VENOUS THROMBOEMBOLISM AFTER SURGERY FOR LONG BONE METASTASES: A RETROSPECTIVE STUDY OF 682 PATIENTS

Olivier Q. Groot, Paul T. Ogink, Stein J. Janssen, Nuno R. Paulino Pereira, Santiago A. Lozano-Calderon, Kevin A Raskin, Francis J. Hornicek, Joseph H. Schwab

*Clinical Orthopaedics and Related Research 2018;476:2052-2061*

# ABSTRACT

### Background

Previous studies have shown that venous thromboembolism (VTE) is a complication associated with neoplastic disease and major orthopaedic surgery. However, many potential risk factors remain undefined.

### Objectives

(1) What proportion of patients develop symptomatic VTE after surgery for long bone metastases?

(2) What factors are associated with the development of symptomatic VTE among patients receiving surgery for long bone metastases?

(3) Is there an association between the development of symptomatic VTE and 1-year survival among patients undergoing surgery for long bone metastases?

(4) Does chemoprophylaxis increase the risk of wound complications among patients undergoing surgery for long bone metastases?

### Design

Retrospective cohort study.

### Methods

A retrospective study identified 682 patients undergoing surgical treatment of long bone metastases between 2002 and 2013 at the Massachusetts General Hospital and Brigham and Women's Hospital. We included patients 18 years of age or older who had a surgical procedure for impending or pathologic metastatic long bone fracture. We considered the humerus, radius, ulna, femur, tibia, and fibula as long bones; metastatic disease was defined as metastases from solid organs, multiple myeloma, or lymphoma. In general, we used 40 mg enoxaparin daily for lower extremity surgery and 325 mg aspirin daily for lower or upper extremity surgery. The primary outcome was a VTE defined as any symptomatic pulmonary embolism (PE) or symptomatic deep vein thrombosis (DVT; proximal and distal) within 90 days of surgery as determined by chart review. The tertiary outcome was defined as any documented wound complication that might be attributable to chemoprophylaxis within 90 days of surgery. At followup after 90 days and 1 year, respectively, 4% (25 of 682) and 8% (53 of 682) were lost to followup. Statistical analysis was performed using multivariable logistic and Cox regression and Kaplan- Meier.

**Results**

Overall, 6% (44 of 682) of patients had symptomatic VTE; 22 patients sustained a DVT, and 22 developed a PE. After controlling for relevant confounding variables, higher preoperative hemoglobin level was independently associated (odds ratio [OR], 0.75; 95% confidence interval [CI], 0.60–0.93; p=0.011) with decreased symptomatic VTE risk, the presence of symptomatic VTE was associated with a worse 1-year survival rate (VTE: 27% [95% CI, 14%–40%] and non-VTE: 39% [95% CI, 35%–43%]; p=0.041), and no association was found between wound complications and the use of chemoprophylaxis (OR, 3.29; 95% CI, 0.43–25.17; p=0.252).

**Conclusion**

The risk of symptomatic 90-day VTE is high in patients undergoing surgery for long bone metastases. Further study would be needed to determine the VTE prevention strategy that best balances risks and benefits to address this complication.

# INTRODUCTION

Venous thromboembolism (VTE), encompassing deep-vein thrombosis (DVT) and pulmonary embolus (PE), is a major public health problem that affects 300,000 to 600,000 individuals in the United States each year and is accompanied by considerable mortality and morbidity.[1–4] The combination of neoplastic disease and major orthopaedic surgery, both of which are known factors associated with VTE[5–13], might put patients at additional risk for developing VTE and could affect survival.

A previous study reported a symptomatic 90-day VTE rate of 10% in 10 of 306 patients undergoing surgery for non-spinal skeletal metastases.[14] However, critical variables in the analysis were absent, including prior local radiotherapy/systemic chemotherapy known for their association with VTE. In addition, this current study examines the relationship between wound complication rate and chemoprophylaxis since this is particularly interesting in the context of the discussion regarding VTE prevention strategies.[15,16] Other studies have also been limited by their sample size, and the fact that they involved heterogeneous patient populations including both primary tumors and metastatic bone lesions.[17–25] Determining factors associated with postoperative VTE development and assessing survival consequences may identify high-risk patients who might benefit from intensified VTE prevention strategies. However, current chemoprophylaxis protocols from national guidelines present ambiguous recommendations about type, dosage and duration after major orthopaedic surgery, let alone after surgery for long bone metastases.[15,16] Accurately balanced chemoprophylaxis protocols are desired to balance between effectively preventing VTE and avoiding wound complications. Determining the relationship between chemoprophylaxis, the rate of symptomatic VTE and wound complications would help clinical decision-making.

We therefore asked (1) What proportion of patients develop symptomatic VTE after surgery for long bone metastases? (2) What factors are associated with the development of symptomatic VTE among patients receiving surgery for long bone metastases? (3) What association exists between the development of symptomatic VTE and one-year survival among patients undergoing surgery for long bone metastases? (4) Does chemoprophylaxis increase the risk of wound complications among patients undergoing surgery for long bone metastases?

# METHODS

### Study design

Our institutional review board approved a waiver of informed consent for this retrospective study at the Massachusetts General Hospital and Brigham and Women's Hospital. The study included 682 patients 18 years of age or older who had a surgical procedure for impending or pathological metastatic long bone fracture between 2002 and 2013. We considered the humerus, radius, ulna, femur, tibia, and fibula as long bones; metastatic disease was defined as metastases from solid organs, multiple myeloma, or lymphoma.[26] The patients included 383 (56%) women and 299 men (44%), with a median age of 64 years (interquartile range [IQR]: 54–72; Table 1).

The median duration of surgery was 191 minutes (IQR: 160–230 minutes) and median hospitalization was 6 days (IQR: 4–9 days). There were 389 (57%) pathological and 293 (43%) impending fractures. Of the 682 fractures, 492 (72%) involved the femur; 160 (23%), the humerus; 25, the tibia; and five, the radius or ulna. Inferior vena cava filters were placed in 17 patients: one (2.3%) in the VTE and 16 (2.5%) in the nonVTE group. Most common primary tumor types included lung (24%), breast cancer (23%) and multiple myeloma (16%) (Table 2).

We excluded patients with: (1) revision procedures, defined as any subsequent procedure after the index surgery addressing the metastatic lesion; (2) surgery due to metastatic fractures in multiple bones; (3) surgical treatment other than intramedullary nailing, plate-screw fixation, endoprosthetic reconstruction, or a combination; and (4) a diagnosed symptomatic VTE within two weeks before surgery since this would interfere with the main aim of the study to find factors associated with developing postoperative VTE. Medical records were flagged with diagnostic and billing codes for prophylactic treatment of an impending fracture or a pathological long bone fracture and then manually checked for eligibility 27. The surgeon selected the operating procedure based on primary tumor type, size, and location of metastatic lesion, estimated survival, and level of disability and pain. Postoperative care and rehabilitation varied based on differences in disease severity.

**Table 1.** Origin of primary tumor (n=682)

| *Primary tumor* | *Number (%)* |
|---|---|
| Lung | 161 (24) |
| Breast | 157 (23) |
| Multiple myeloma | 109 (16) |
| Kidney | 54 (7.9) |
| Lymphoma | 37 (5.4) |
| Prostate | 31 (4.6) |
| Melanoma | 22 (3.2) |
| Thyroid | 15 (2.2) |
| Esophageal | 14 (2.1) |
| Colorectal | 12 (1.8) |
| Hepatocellular | 10 (1.5) |
| Adenocarcinoma of unknown origin | 9 (1.3) |
| Other* | 51 (7.5) |

** This category included 10 patients with an unknown cancer, seven with bladder cancer, five with ovarian cancer, five with neuroendocrine cancer, five with skin cancer, four with pancreatic cancer, four with salivary gland cancer, three with endometrial cancer, three with hemangioendothelioma, two with vulvar cancer, two with blue-cell tumor, and one patient with gastric cancer.*

### Outcomes and Explanatory Variables

We obtained data through chart review by two independent research fellows. Our primary outcome was a symptomatic VTE, presenting with swelling, redness or pain of the lower extremities or problems with breathing, defined as any symptomatic pulmonary embolism (PE) or symptomatic distal or proximal deep vein thrombosis (DVT) within 90 days of surgery diagnosed with the following diagnostic procedures: venography, impedance plethysmography, pulmonary arteriography, chest CT, ventilation-perfusion lung scan, and vascular ultrasound. Our secondary outcome was survival after surgery. October 1, 2016 was considered as the final date of follow-up for survival outcome assessment. We determined date of death by using the Social Security Index and medical charts. At follow-up after 90 days and one-year, respectively 4% (25/682) and 8% (53/682) were lost to follow-up. Our third outcome was documented wound complications, defined as a wound complication that might be attributable to chemoprophylaxis within 90 days of surgery, categorized in: nine deep infections treated with irrigation and debridement, five superficial wound complications consisting of three wound dehiscences that were treated surgically, and two hematomas that were treated without surgery, and four deep wound complications consisting of three hematomas and one retroperitoneal bleed treated surgically.[28] Wound complications such as wound inflammation requiring antibiotics were disregarded. Only two patients had a wound complication followed by a symptomatic VTE.

During the period in question, we generally used either enoxaparin 40 mg or aspirin 325 mg daily for patients operated on the lower extremity. For surgery on the upper extremity, we used aspirin

**Table 2.** Patient and treatment characteristics for the no VTE and VTE group (n=682)

| *Variables* | *No VTE (n=638)* | *VTE (n=44)* |
|---|---|---|
| | *Median (IQR)* | |
| Age (years) | 64 (54-72) | 62 (56-69) |
| Modified Charlson Comorbidity Index[a] | 6 (6-8) | 6 (6-7) |
| Total estimated blood loss during surgery (mL)[b] | 200 (100-400) | 200 (100-300) |
| Duration of surgery (minutes)[b] | 191 (160-230) | 200 (158-241) |
| Duration of primary diagnosis until metastatic operation (days) | 617 (77-2078) | 200 (27-2794) |
| Duration of hospitalization (days) | 6 (4-9) | 7 (5-15) |
| Total perioperative transfused[c] | 1 (0-2) | 1 (0-2) |
| Preoperative laboratory values[b] | | |
| Hemoglobin levels (g/dL) | 11 (10-12) | 10 (9-11) |
| White blood cell count (10$^3$/ L) | 9 (6-13) | 9 (6-13) |
| Creatinine levels (mg/dL) | 0.8 (0.6-1.0) | 0.7 (0.6-0.9) |
| Calcium levels (mg/dL) | 9 (8-9) | 9 (8-9) |
| Platelet count (10$^3$/mm$^3$) | 237 (181-311) | 259 (188-343) |
| | *Number (%)* | |
| Men | 285 (45) | 14 (32) |
| Body mass index (in kg/m$^2$) | | |
| < 18.5 | 19 (3.3) | 1 (2.2) |
| 18.5-30 | 409 (72) | 25 (57) |
| > 30 | 142 (25) | 14 (32) |
| Smoking status | | |
| Never smoked | 244 (39) | 16 (36) |
| Former smoker | 284 (45) | 16 (36) |
| Current smoker | 100 (16) | 11 (25) |
| Pathologic fracture | 365 (57) | 24 (55) |
| Type of surgery | | |
| Intramedullary nail | 394 (62) | 31 (70) |
| Endoprosthetic reconstruction | 129 (20) | 9 (20) |
| Plate and screw fixation | 115 (18) | 4 (9.1) |
| Metastases region[d] | | |
| Lower extremities | 480 (75) | 37 (84) |
| Upper extremities | 158 (25) | 7 (16) |
| Multiple bone metastases[e] | 485 (76) | 36 (82) |
| Visceral metastases | 291 (46) | 22 (50) |
| Prior embolization | 20 (3.1) | 1 (2.3) |
| IVC filter prophylaxis | 16 (2.5) | 1 (2.3) |
| Previous local radiotherapy | 120 (19) | 7 (16) |
| Previous systemic therapy | 403 (63) | 26 (59) |

*VTE=venous thromboembolism; IQR=interquartile range; IVC=inferior vena cava.*

*a These values were based on any additional comorbidity in addition to the metastatic disease score according to the modified Charlson Comorbidity Index.*

*b Estimated blood loss was available in 583 patients (91%) from the no VTE group and in 38 patients (86%) from the VTE group; duration of surgery in 636 patients (100%) from the no VTE group and in 44 patients (100%) from the VTE group; preoperative hemoglobin level in 595 patients (93%) from the no VTE group and in 41 patients (93%) from the VTE group; preoperative white blood cell count in 604 patients (95%) from the no VTE group and in 42 patients (95%) from the VTE group; preoperative creatinine levels in 575 patients (90%) from the no VTE group and in 41 patients (93%) from the VTE group; preoperative calcium level in 485 patients (76%) from the no VTE group and in 38 patients (86%) from the VTE group; preoperative platelet count in 603 patients (95%) from the no VTE group and in 42 patients (95%) from the VTE group; body mass index in 570 patients (89%) from the no VTE group and in 40 patients (91%) from the VTE group; and smoking status in 628 patients (98%) from the no VTE group and in 43 patients (98%) from the VTE group.*

*c Total perioperative transfused includes all blood and nonblood products.*

*d Lower extremities included 492 (72%) femurs and 25 (3.7%) tibiae. Upper extremity included 160 (23%) humerus, three (0.4%) radius, and two (0.3%) ulnae.*

*e Any bone metastasis outside of the treated lesion.*

325 mg daily for major reconstruction and no chemoprophylaxis for less invasive surgery. Other general thromboembolic prophylactic dosages used were: dalteparin 5000 IUs daily, warfarin dependent to maintain an international normalized ratio of 2.0:2.5, and subcutaneous heparin 5000 IUs every 12 hours. Patients on preoperative chemoprophylaxis continued their initial medication postoperatively. All chemoprophylaxis was started 6-12 hours after surgery and continued day-to-day but was discontinued if a bleeding complication developed. In case of contra-indications for chemoprophylaxis an IVC filter was placed before surgery on the lower extremity and no chemoprophylaxis was prescribed for surgery on the upper extremity. A total of 17 IVC filters were placed. Chemical anticoagulants, within a maximum range of 14 days postoperative, were considered prophylactic. The most-aggressive chemoprophylaxis regimen was considered in our analyses in case of overlapping regimens. The following anticoagulant regimens were used: low-molecular-weight heparin (LMWH) for 358 of 682 patients (52%); no form of chemical anticoagulant for 113 patients (17%); warfarin for 129 patients (19%); aspirin for 66 (10%) patients; subcutaneous heparin for 16 (2%). Compression stockings and sequential compression devices were not included as potential variables because they were routinely employed as mechanical prophylaxis at both centers in all patients after surgery throughout their hospitalization.

Preoperative laboratory values, nearest to surgery with a maximum range of 7 days, included: hemoglobin level (g/dL), creatinine level (mg/dL), calcium level (mg/dL), white blood cell count (1000/mm3), and platelet count (1000/mm3). Fracture type was defined as pathological or impending fracture. Impending fractures were considered imminent to pathologic fracture if they possessed a destructive bone lesion with no visible fracture line, loss of height, rotation, or angulation. The surgeon determined operative treatment, based on the severity of pain and the degree of destruction, to prevent a pathological fracture. The patient comorbidity status was determined using the modified Charlson Comorbidity Index.[29,30] An ICD-9 code-based algorithm classifying 12 comorbidities preoperatively provided a score ranging from 0 to 24[30], with a higher score corresponding with a more severe comorbidity status.[27] Placement of an inferior vena cava filter was considered prophylactic when it was placed preoperatively or within 90 days postoperatively. Inferior vena cava filters placed after VTE were disregarded. We determined operative treatment time in minutes using the anesthesia time as a surrogate marker, which measured the presence of the patient in the operating room from arrival until departure.

### Statistical Analysis

We used multivariable logistic regression analysis, controlling for confounding variables identified in bivariate testing with a p value $< 0.10$ and presumed to be relevant to VTE[14,31], to assess independent risk factors for symptomatic VTE. Odds ratio for continuous variables are interpreted in terms of each unit increase or decrease on the scale (i.e., 1 to 2, 2 to 3, etc.). Bivariate analysis found that higher blood loss during surgery (OR, 1.00; 95% CI, 1.00–1.00; p=0.036) and higher preoperative hemoglobin levels (OR, 0.74; 95% CI, 0.60–0.92; p=0.007) were associated with decreased, and longer duration of hospitalization (OR, 1.06; 95% CI, 1.02–1.10; p=0.003) with increased risk of symptomatic VTE development (Appendix 1). Lung cancer (OR, 1.74; 95% CI, 0.91–3.34; p=0.094) was included in the multivariable analysis due to a P-value of < 0.1. Additional variables controlled for, were age, the modified Charlson Comorbidity Index, visceral metastases, and chemoprophylaxis. Multivariate logistic regression was also used to assess the relation between chemoprophylaxis and wound complications, controlling for age and the modified Charlson Comorbidity Index. We used Cox regression analysis, after controlling for the confounding factors age, gender, BMI, the modified Charlson Comorbidity Index, visceral and other bone metastases, estimated blood loss, operation type, and pathological fracture, to determine differences in survival between the symptomatic VTE and nonVTE group. Kaplan-Meier plots demonstrated the survival curves for both groups. We applied multiple imputations to estimate missing values for estimated blood loss during surgery (61 of 682 patients) and preoperative hemoglobin levels (46 of 682 patients). A two-tailed P-value < 0.05 was considered significant. All statistical analyses were performed using Stata 13.0 (StataCorp LP, College Station, TX, USA).

# RESULTS

Symptomatic VTE was diagnosed in 6% (44 of 682) patients; 22 had a PE and 22 had a DVT (Table 3). The median age of the 44 patients was 62 years (IQR: 56–69 years), and 14 (32%) were men. Of the 22 patients with a PE, two had a confirmed DVT one day later, seven tested negative on CT or ultrasound, and 13 did not undergo assessment of DVT presence. One patient died 5 days postoperative due to PE. Symptomatic VTE was diagnosed in six patients with metastatic multiple myeloma and two patients with lymphoma. More than half of the patients (57%) developed a symptomatic VTE after their postoperative discharge; the median postoperative hospitalization of these patients was 7 days (IQR: 5–15 days), and the median time between surgery and symptomatic VTE development was 12 days (IQR: 4–45 days, last symptomatic VTE event documented at 85 days after surgery).

After controlling for potentially relevant confounding variables such as age and lung-cancer histology, we found that the following two factors were independently associated with respectively increased and decreased risk of symptomatic VTE development: longer duration of hospitalization (OR, 1.06; 95% CI, 1.02–1.11; p=0.006), and higher preoperative hemoglobin levels (OR, 0.75; 95% CI, 0.60–0.93; p=0.011; Table 4). Symptomatic VTE occurred in 39 of 569 patients for chemoprophylaxis in its entirety and in five of 113 patients for no chemoprophylaxis, demonstrating no association after controlling for age, gender, the modified Charlson Comorbidity Index, and lung-cancer histology (OR, 1.65; 95% CI, 0.63–4.32; p=0.310). Patients who had a symptomatic VTE within 90 days had lower 1-year survival than did those without symptomatic VTE, after controlling for the confounding variables; age, gender, BMI, fracture the modified Charlson Comorbidity Index comorbidity index, visceral and other bone metastases, estimated blood loss, operation type and pathological fracture (27% [95% CI: 14%– 40%] and 39% [95% CI: 35%–43%; p=0.041]) (Figure 1). The probability of developing a symptomatic VTE rose gradually with a notable increase at 30 days after surgery (Figure 2). Timing of symptomatic VTE ranged from day 1 to day 85.

With the numbers available, we found no association, after controlling for age and the modified Charlson Comorbidity Index, between any of the studied chemoprophylaxis regimens and the occurrence of 18 wound complications, consisting of 10 (56%) for LMWH (reference value), 6 (33%) for warfarin (OR, 1.40; 95% CI, 0.47 – 4.19; p=0.547), 1 (6%) for aspirin (OR, 0.54; 95% CI, 0.07 – 4.32; p=0.563), 1 (6%) for no form of chemoprophylaxis (OR, 0.31; 95% CI, 0.04 – 2.43; p=0.263), and 0 for heparin (no values available). An additional subanalysis between chemoprophylaxis in its entirety and no chemoprophylaxis also showed no difference, but this was underpowered.

**Table 3.** 90-Day symptomatic VTE, wound complications and anticoagulant use (n=682)

| *Variables* | *All patients (n=682)* |
|---|---|
| | *Number (%)* |
| VTE events | 44 (6.5) |
| PE | 22 (3.2) |
| DVT | 22 (3.2) |
| Wound complications | 18 (2.6) |
| Superficial infections | 0 (0) |
| Deep infections | 9 (1.3) |
| Superficial wound complications | 5 (0.7) |
| Deep wound complications | 4 (0.6) |
| Anticoagulant | |
| None | 113 (17) |
| Low molecular weight heparin | 358 (52) |
| Warfarin | 129 (19) |
| Aspirin | 66 (9.7) |
| Subcutaneous heparin | 16 (2.3) |
| | *Median (IQR)* |
| Time between surgery and VTE (days) | 12 (4-45) |
| PE | 7 (4-41) |
| DVT | 24 (7-51) |
| Duration postoperative hospitalization for VTE patients (days)* | 7 (5-15) |
| PE | 6 (5-13) |
| DVT | 8 (5-16) |

*VTE=venous thromboembolism; PE=pulmonary embolism; DVT=deep vein thrombosis; IQR=interquartile range.*
**Time in days between operation and discharge for patients with VTE; the VTE developed during or after this period.*

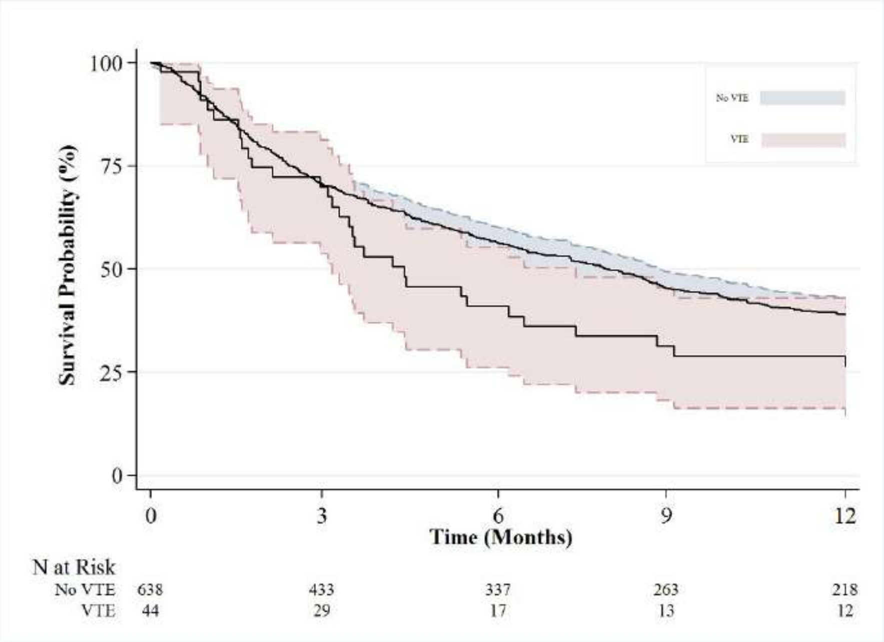


**Figure 1.** Kaplan-Meier plot demonstrating the survival probability with 95% CIs for patients with and without postoperative symptomatic VTE (p=0.041).

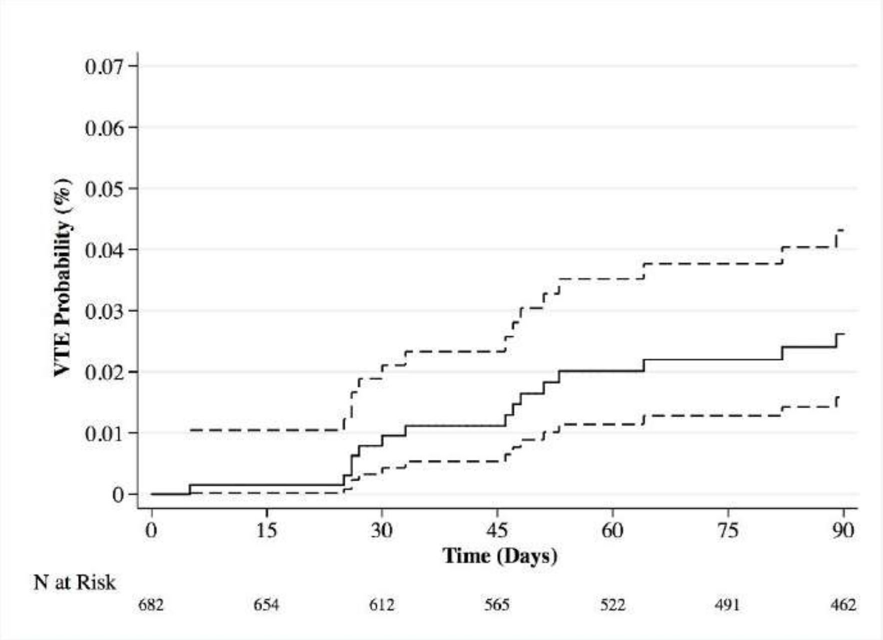


**Figure 2.** . Kaplan-Meier plot demonstrating the probability of developing a symptomatic VTE within 90 days postoperatively 0.03% (95% CI, 0.02–0.04). The risk of symptomatic VTE occurrence has a sudden increase at 30 days postoperatively and keeps increasing gradually thereafter.

**Table 4.** Multivariate logistic regression assessing risk factors for 90-day VTE after multiple imputation (40 imputations) (n=682)

| *Variables* | *Odds ratio (95% CI)* | *Standard error* | *P-value* |
|---|---|---|---|
| Total estimated blood loss during surgery (mL)* | 1.00 (1.00 – 1.00) | 0.00 | 0.111 |
| Lung cancer | 1.71 (0.86 – 3.41) | 0.60 | 0.127 |
| Duration hospitalization (days) | 1.06 (1.02 – 1.11) | 0.02 | **0.006** |
| Preoperative hemoglobin levels (g/dL)* | 0.75 (0.60 – 0.93) | 0.09 | **0.011** |
| Age (years) | 1.00 (0.97 – 1.03) | 0.01 | 0.997 |
| Modified Charlson Comorbidity Index | 0.87 (0.74 – 1.02) | 0.07 | 0.080 |
| Visceral metastases | 1.08 (0.55 – 2.14) | 0.38 | 0.816 |
| Anticoagulant | | | |
| None | 0.51 (0.18 – 1.41) | 0.26 | 0.193 |
| Low molecular weight heparin | *Reference value* | | |
| Subcutaneous heparin | 0.87 (0.11 – 7.01) | 0.92 | 0.894 |
| Aspirin | 0.47 (0.10 – 2.08) | 0.36 | 0.317 |
| Warfarin | 1.07 (0.46 – 2.48) | 0.46 | 0.883 |

*VTE=venous thromboembolism; CI=confidence interval.* Bold *indicates significance (P<0.05).*
**Estimated blood loss was available in 583 patients (91%) from the no VTE group and in 38 patients (86%) from the VTE group; preoperative hemoglobin level was available in 595 patients (93%) from the no VTE group and in 41 patients (93%) from the VTE group*

# DISCUSSION

Patients undergoing major surgery for bone metastases are at high risk for developing postoperative VTE. This study is advantaged over prior work considering the more than double sample of patients with long bone metastases and extensive follow-up considering the cohort's clinical characteristics adding more than 15 potential variables for analysis. The symptomatic VTE incidence was 6% and after controlling for potential confounding variables such as age and lung-cancer histology, we found that preoperative hemoglobin levels and duration of hospitalization were independently associated with symptomatic VTE development. After controlling for confounding variables, patients with symptomatic VTE had worse survival, though only one of the 682 (0.1%) patients died from PE. There was no association between any of the studied anticoagulant regimens and the development of symptomatic VTE or postoperative wound complications. However, we were not sufficiently powered to address this problem.

This study has limitations. First, because of the retrospective study design there was no uniform anticoagulant regimen or standardization regarding chemoprophylaxis use. We used in general enoxaparin 40mg daily for lower extremity surgery and aspirin 325mg daily for lower or upper surgery. Second, uncontrolled for differences likely exist in the survival analysis between the symptomatic VTE and non-VTE group, since only one fatal PE was identified. However, we controlled for multiple confounding survival variables such as age and the modified Charlson Comorbidity Index. Third, we could not confirm the exact duration and compliance of anticoagulant use for all patients because chemoprophylaxis duration during and after hospitalization was not recorded and post-discharge compliance was not monitored. However, both institutions maintained a protocol that required patients to use anticoagulants for four weeks postoperatively. Additionally, it was protocol that patients use sequential compression devices and compression stockings during postoperative hospitalization. Fourth, we included only symptomatic VTE because of the lack of a screening protocol, which likely resulted in an underestimated VTE incidence. We anticipate a relatively low number of clinically relevant VTEs were missed as this complicated patient population was closely monitored by healthcare providers and frequently visited the clinic postoperatively. Lastly, metastases from lymphoma and multiple myeloma were included, which are known for their increased symptomatic VTE risk and better prognosis.[32] Nonetheless, we included them because these metastases represent 21% (146/682) of the long bone metastases.

In this series, 6% of patients (44 of 682) developed symptomatic VTE and 3% (22 of 682) developed PE. This symptomatic 90-day VTE rate is within the reported symptomatic VTE range of 2.7% to 28% of comparable musculoskeletal metastases series.[14,17–23,25,31] The Ratasvouri paper reported comparable results of symptomatic 90-day VTE rate of 10% (30 of 306), PE of 3.3% (10/306), poor survival of patients with VTE and the late-onset of postoperative VTE development.[14] In addition, higher preoperative hemoglobin was identified as a factor associated with decreased postoperative symptomatic VTE development and more than 15 potential variables were added to the analysis. Also, although underpowered, this study elaborated on the discussion regarding wound complication and chemoprophylaxis in the context of VTE prevention strategies.

Although the proportion of symptomatic VTE was considerable, for fatal PE it is low, occurring in only one of 682 patients, meaning that these patients are not dying as a direct cause of symptomatic VTE. Thrombocytosis is an independent predictor of survival in multiple cancers[33,34] via enhanced invasiveness of tumor cells[35], promotion of tumor cell motility36 and stimulation of epithelial-mesenchymal transition.[37] However, the mechanism by which symptomatic VTE and nonfatal PE are associated with worse survival is poorly understood. A recent study demonstrated that activated platelets, which are seen during VTE, may inhibit the immune response to cancer cells by facilitating T lymphocyte inhibition through binding to transforming growth factor-β (TGF- β) in serum. The study further postulated that antiplatelet therapy may be an effective adjunct to immunotherapy.[38] This data should be considered when considering chemoprophylaxis in the setting of surgery for metastatic disease. Another possible reason for the poor survival in VTE patients is the highly complex patient population, with multiple comorbidities and other disease-related factors, where patients with more-advanced cancer develop VTE more easily. Moreover, only one fatal PE was confirmed indicating that VTE may function more as a predictor than as the main cause for poor survival.

The lower extremity, and especially the femur, was the most common surgery site. Although lower extremity procedures have a greater effect on patient mobility, the location of surgery was not a factor associated with symptomatic VTE development in this study. It is possible that the risk caused by skeletal metastatic disease–malignancy promotes a hypercoagulable state–substantially outweighs the risk of immobility due to lower extremity surgery. A higher preoperative hemoglobin level was an independent factor associated with decreased symptomatic VTE. This biomarker was previously identified as a predictive marker for cancer-associated VTE.[39] Clinically, preoperative hemoglobin should be incorporated into the risk adjustment as a factor associated with postoperative symptomatic VTE.

Duration of hospitalization was also independently but only slightly associated with increased symptomatic VTE. We note that this association between longer duration of hospitalization and development of symptomatic VTE does not imply that longer hospitalization causes symptomatic VTE. It may well be the other way around, the occurrence of symptomatic VTE results in longer hospitalization. This is demonstrated by comparing different means of hospitalization. The means of hospitalization for all non-VTE patients (7 ± 6 days) and symptomatic VTE patients that developed during hospitalization (14 ± 6 days) are quite different, but this longer duration of hospitalization is preceded by early development of symptomatic VTE in this group (4 ± 3 days). Additionally, the means of hospitalization for all nonVTE patients (7 ± 6 days) and symptomatic VTE patients that developed after discharge (8 ± 7 days) are nearly identical, resulting in no difference (Appendix 2). This indicates that the occurrence of symptomatic VTE during hospitalization may result in a longer hospitalization and that longer duration of hospitalization should not be considered as a factor associated with symptomatic VTE development.

No association was found between a specific anticoagulant regimen and the development of wound complications. However, this analysis was underpowered to detect a relationship given the relatively low rate of wound complications within each separate anticoagulant group. Shallop et al.31 reported a comparably low risk of wound complications and infection in intramedullary nailing for metastatic bone lesions, as did a systematic review including 3211 metastatic lesions in the femur.40 The risk of major wound complications seems low, with only nine of 682 patients undergoing revision surgery for deep infection, three for a deep, large hematoma, and one patient who developed a large retroperitoneal hematoma and was admitted to the intensive care unit. Meanwhile, symptomatic VTE development is considerable in this population. Considering these results, future studies need to determine the relation between wound complications and various anticoagulant agents, as well as the ideal prophylactic dosage to address the high rate of symptomatic VTE.[31]

The probability of developing a symptomatic VTE increased precipitously 30 days postoperatively and steadily increased over a 90-day period (Figure 2). Previous studies have shown that the risk of postoperative symptomatic VTE persists for several weeks after hospital discharge in patients undergoing high-risk orthopaedic surgery.[41–44] Correspondingly, most symptomatic VTEs, which occurred in 25 of 44 patients (57%), developed after discharge; for patients with symptomatic VTE the time in days between surgery and symptomatic VTE (median 12 days; IQR 4–45 days) was considerably longer than the duration of postoperative hospitalization (median 7 days, IQR 5–15 days). The onset of late postoperative symptomatic VTE was also observed in comparable studies, and it has been suggested longer duration of prophylaxis may prevent this.[31,45] However, most major national orthopaedic guidelines remain unclear about the duration of anticoagulant usage, stating that patients and physicians should discuss the duration of prophylaxis.[16] Although this study was not designed to specifically address this compliance variable, previous studies report poor compliance of outpatient anticoagulant usage and inappropriate prophylaxis prescription at discharge.[46–48] Moreover, given the tendency toward shorter hospitalization after major orthopaedic surgery, the importance of compliance of outpatient anticoagulants in preventing postoperative symptomatic VTE must be stressed.[49] Novel oral anticoagulants could fulfill a prominent role, since most patients prefer oral agents.[50] Further study should elucidate the ideal duration of postoperatively prophylactic regimen. Interestingly, 10 of the 22 symptomatic DVT patients (45%) had a DVT away from the site of surgery; five patients (23%) had DVTs isolated to the contralateral limb and five patients (23%) had bilateral DVTs. This supports the concept that systemic factors stimulate thrombosis in cancer patients in addition to local and mechanical factors.

# CONCLUSION

This study presents a high symptomatic 90-day VTE rate among patients undergoing surgery for long bone metastases, warranting several considerations. First, protocols may need to incorporate patient-specific risk factors, such as preoperative hemoglobin levels. Second, future studies should elucidate the ideal postoperative VTE prevention regimen. In concordance with similar studies, the risk for VTE is clearly high, requiring further investigation.

# REFERENCES

1. Heit JA, Silverstein MD, Mohr DN, et al. Risk factors for deep vein thrombosis and pulmonary embolism: a population-based case-control study. *Arch Intern Med.* 2000;160(6):809–15.

2. Silverstein MD, Heit JA, Mohr DN, et al. Trends in the incidence of deep vein thrombosis and pulmonary embolism. *Arch Intern Med.* 1998;158(6):585.

3. Spencer FA, Emery C, Lessard D, et al. The Worcester venous thromboembolism study: A population-based study of the clinical epidemiology of venous thromboembolism. *J Gen Intern Med.* 2006;21(7):722–727.

4. White RH, Zhou H, Murin S, et al. Effect of ethnicity and gender on the incidence of venous thromboembolism in a diverse population in California in 1996. *Thromb Haemost.* 2005;93(2):298–305.

5. Amato A, Pescatori M. Perioperative blood transfusions and recurrence of colorectal cancer. *Cochrane Database of Systematic Reviews.* 2006:CD005033.

6. Behranwala KA, Williamson RC. Cancer-associated venous thrombosis in the surgical setting. *Ann Surg.* 2009;249(3):366–375.

7. Clagett GP, Anderson Jr. FA, Geerts W, et al. Prevention of venous thromboembolism. *Chest.* 1998;114(5 Suppl):531S-560S.

8. Geerts WH, Bergqvist D, Pineo GF, et al. Prevention of venous thromboembolism: American College of Chest Physicians evidence-based clinical practice guidelines (8th edition). *Chest.* 2008;133(6 SUPPL. 6):381S-453S.

9. Karadimas EJ, Papadimitriou G, Theodoratos G, et al. The effectiveness of the antegrade reamed technique: The experience and complications from 415 traumatic femoral shaft fractures. *Strateg Trauma Limb Reconstr.* 2009;4(3):113–121.

10. McLaughlin DF, Wade CE, Champion HR, et al. Thromboembolic complications following trauma. *Transfusion.* 2009;49(SUPPL.5):256S-263S.

11. Montgomery KD, Geerts WH, Potter HG, et al. Thromboembolic complications in patients with pelvic trauma. *Clin Orthop Relat Res.* 1996;329(329):68–87.

12. Owings JTJ, Gosselin R. Acquired antithrombin deficiency following severe traumatic injury: rationale for study of antithrombin supplementation. *Semin Thromb Hemost.* 1997;23(suppl 1):17–24.

13. Planès A, Vochelle N, Fagola M. Total hip replacement and deep vein thrombosis. A venographic and necropsy study. *J Bone Joint Surg Br.* 1990;72(1):9–13.

14. Ratasvuori M, Lassila R, Laitinen M. Venous thromboembolism after surgical treatment of non-spinal skeletal metastases - An underdiagnosed complication. *Thromb Res.* 2016;141:124–8.

15. Falck-Ytter Y, Francis CW, Johanson NA, et al. Prevention of VTE in orthopedic surgery patients. Antithrombotic therapy and prevention of thrombosis, 9th ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest.* 2012;141:e278S-e325S.

16. Mont MA Jacobs JJ et al. Preventing venous thromboembolic disease in patients undergoing elective hip and knee arthroplasty guideline. *J Am Acad Orthop Surg.* 2011;19(Dec):768–776.

17. Benevenia J, Bibbo C, Patel DV, et al. Inferior vena cava filters prevent pulmonary emboli in patients with metastatic pathologic fractures of the lower extremity. *Clin Orthop Relat Res.* 2004;(426):87–91.

18. Damron TA, Wardak Z, Glodny B, et al. Risk of venous thromboembolism in bone and soft-tissue sarcoma

patients undergoing surgical intervention: A report from prior to the initiation of SCIP measures. *J Surg Oncol.* 2011;103(7):643–647.

19. Lin PP, Graham D, Hann LE, et al. Deep venous thrombosis after orthopedic surgery in adult cancer patients. *J Surg Oncol.* 1998;68(1):41–7.

20. Mitchell SY. Venous thromboembolism in patients with primary bone or soft-tissue sarcomas. *J Bone Jt Surg.* 2007;89(11):2433.

21. Morii T, Mochizuki K, Tajima T, et al. Venous thromboembolism in the management of patients with musculoskeletal tumor. *J Orthop Sci.* 2010;15(6):810–815.

22. Nathan SS, Simmons KA, Lin PP, et al. Proximal deep vein thrombosis after hip replacement for oncologic indications. *J Bone Joint Surg Am.* 2006;88(5):1066–70.

23. Ogura K, Yasunaga H, Horiguchi H, et al. Incidence and risk factors for pulmonary embolism after primary musculoskeletal tumor surgery. *Clin Orthop Relat Res.* 2013;471(10):3310–3316.

24. Patel AR, Crist MK, Nemitz J, et al. Aspirin and compression devices versus low-molecular-weight heparin and PCD for VTE prophylaxis in orthopedic oncology patients. *J Surg Oncol.* 2010;102(3):276–281.

25. Tuy B, Bhate C, Beebe K, et al. IVC filters may prevent fatal pulmonary embolism in musculoskeletal tumor surgery. *Clin Orthop Relat Res.* 2009;467(1):239–245.

26. Nathan SS, Healey JH, Mellano D, et al. Survival in patients operated on for pathologic fracture: Implications for end-of-life orthopedic care. *J Clin Oncol.* 2005;23(25):6072–6082.

27. Janssen SJ, Braun Y, Ready JE, et al. Are allogeneic blood transfusions associated with decreased survival after surgery for long-bone metastatic fractures? *Clin Orthop Relat Res.* 2015;473(7):2343–2351.

28. Ramo BA, Griffin AM, Gill CS, et al. Incidence of symptomatic venous thromboembolism in oncologic patients undergoing lower-extremity endoprosthetic arthroplasty. *J Bone Joint Surg Am.* 2011;93(9):847–54.

29. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J Chronic Dis.* 1987;40(5):373–383.

30. Quan H, Li B, Couris CM, et al. Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol.* 2011;173(6):676–682.

31. Shallop B, Starks A, Greenbaum S, et al. Thromboembolism after intramedullary nailing for metastatic bone lesions. *J Bone Jt Surgery-American Vol.* 2015;97(18):1503–1511.

32. Katagiri H, Okada R, Takagi T, et al. New prognostic factors and scoring system for patients with skeletal metastasis. *Cancer Med.* 2014;3(5):1359–1367.

33. Sierko E, Wojtukiewicz MZ. Platelets and angiogenesis in malignancy. *Semin Thromb Hemost.* 2004;30(1):95–108.

34. Stone JH, Zen Y, Deshpande V. IgG4-Related Disease. *N Engl J Med.* 2012;366(6):539–551.

35. Karpatkin S, Pearlstein E, Ambrogio C, et al. Role of adhesive proteins in platelet tumor interaction in vitro and metastasis formation in vivo. *J Clin Invest.* 1988;81(4):1012–1019.

36. Schumacher D, Strilic B, Sivaraj K, et al. Platelet-Derived Nucleotides Promote Tumor-Cell Transendothelial Migration and Metastasis via P2Y2 Receptor. *Cancer Cell.* 2013;24(1):130–137.

37. Labelle M, Begum S, Hynes RO. Direct signaling between platelets and cancer cells induces an epithelial-

**mesenchymal-like transition and promotes metastasis.** *Cancer Cell.* 2011;20(5):576–590.

38. Rachidi S, Metelli A, Riesenberg B, et al. **Platelets subvert T cell immunity against cancer via GARP-TGF⊠ axis.** *Sci Immunol.* 2017;2(11):eaai7911.

39. Khorana A, Kuderer NM, Culakova E, et al. **Development and validation of a predictive model for chemotherapy- associated thrombosis.** *Blood.* 2008;111(10):4902–4907.

40. Janssen SJ, Teunis T, Hornicek FJ, et al. **Outcome after fixation of metastatic proximal femoral fractures: A systematic review of 40 studies.** *J Surg Oncol.* 2016;114(4):507–519.

41. Bergqvist D. **The postdischarge risk of venous thromboembolism after hip replacement. The role of prolonged prophylaxis.** *Drugs.* 1996;52 Suppl 7:55–59.

42. Bergqvist D, Agnelli G, Cohen AT, et al. **Duration of prophylaxis against venous thromboembolism with enoxaparin after surgery for cancer.** *N Engl J Med.* 2002;346(13):975–980.

43. Bergqvist D, Lindblad B. **A 30-year survey of pulmonary embolism verified at autopsy: an analysis of 1274 surgical patients.** *Br J Surg.* 1985;72(2):105–108.

44. Warwick D, Friedman RJ, Agnelli G, et al. **Insufficient duration of venous thromboembolism prophylaxis after total hip or knee replacement when compared with the time course of thromboembolic events: findings from the Global Orthopaedic Registry.** *J Bone Joint Surg Br.* 2007;89(6):799–807.

45. Sweetland S, Green J, Liu B, et al. **Duration and magnitude of the postoperative risk of venous thromboembolism in middle aged women: prospective cohort study.** *BMJ.* 2009;339(dec03 1):b4583–b4583.

46. Bergqvist D, Arcelus JI, Felicissimo P. **Evaluation of the duration of thromboembolic prophylaxis after high-risk orthopaedic surgery: The ETHOS observational study.** *Thromb Haemost.* 2012;107(2):270–279.

47. Gao Y, Long A, Xie Z, et al. **The compliance of thromboprophylaxis affects the risk of venous thromboembolism in patients undergoing hip fracture surgery.** *Springerplus.* 2016;5(1):1362.

48. Wilke T, Müller S. **Nonadherence in outpatient thromboprophylaxis after major orthopedic surgery: a systematic review.** *Expert Rev Pharmacoecon Outcomes Res.* 2010;10(6):691–700.

49. OECD. **OECD Health Data 2009 – comparing health statistics across OECD countries - OECD.** Available at: http://www.oecd.org/health/oecdhealthdata2009comparinghealthstatisticsacrossoecdcountries.htm. Accessed January 15, 2018.

50. Friedman RJ. **Novel Oral Anticoagulants for VTE Prevention in Orthopedic Surgery: Overview of Phase 3 Trials.** *Orthopedics.* 2011;34(10):795–804.

# SUPPLEMENTAL MATERIAL TO CHAPTER 7

**Appendix 1.** Bivariate logistic regression assessing risk factors for 90-day VTE events (n=682)

**Appendix 2.** Comparison of hospitalization in days between patients with and without venous thromboembolism.

Supplemental material can be consulted online per the website of the journal and/or publisher.

CHAPTER

08

# HIGH RISK OF SYMPTOMATIC VENOUS THROMBOEMBOLISM AFTER SURGERY FOR SPINE METASTATIC BONE LESIONS: A RETROSPECTIVE STUDY

Olivier Q. Groot, Paul T. Ogink, Nuno R. Paulino Pereira, Marco L. Ferrone, Mitchell B. Harris, Santiago A. Lozano-Calderon, Andrew J. Schoenfeld, Joseph H. Schwab

*Clinical Orthopaedics and Related Research 2019;477:1674-1686*

# ABSTRACT

### Background

Cancer and spinal surgery are both considered risk factors for venous thromboembolism (VTE). However, the risk of symptomatic VTE for patients undergoing surgery for spine metastases remains undefined.

### Objectives

(1) Identify the proportion of patients who develop symptomatic VTE within 90-days of surgical treatment for spine metastases

(2) Identify the factors associated with the development of symptomatic VTE among patients receiving surgery for spine metastases.

(3) Assess the association between the development of postoperative symptomatic VTE and 1-year survival among patients who underwent surgery for spine metastases

(4) Assess if chemoprophylaxis increases the risk of wound complications among patients who underwent surgery for spine metastases.

### Design

Retrospective cohort study.

### Methods

Between 2002 and 2014, 637 patients at two hospitals underwent spine surgery for metastases. We considered eligible for analysis adult patients whose procedures were to treat cervical, thoracic, or lumbar metastases (including lymphoma and multiple myeloma). At followup after 90 days and 1 year, respectively, 21 of 637 patients (3%) and 41 of 637 patients (6%) were lost to followup. In general, we used 40 mg of enoxaparin or 5000 IUs subcutaneous heparin every 12 hours. Patients on preoperative chemoprophylaxis continued their initial medication postoperatively. All chemoprophylaxis was started 48 hours after surgery and continued day to day but was discontinued if a bleeding complication developed. Low-molecular-weight heparin (including enoxaparin and dalteparin, in general dosages of respectively 40 mg and 5000 IUs daily) was the most used chemoprophylaxis in 308 patients (48%). Subcutaneous heparin was injected into 127 patients (20%); aspirin was used for 92 patients (14%); and warfarin was administered in 21 patients (3.3%). No form of chemoprophylaxis was prescribed for 89 patients (14%). The primary outcome variable, VTE, was defined as any symptomatic pulmonary embolism (PE) or symptomatic deep venous thromboembolism (DVT) within 90 days of surgery as determined by chart review. The secondary outcome was defined as any documented wound complication within 90 days of surgery that might be attributable to chemoprophylaxis. Statistical analysis was performed using multivariable logistic and Cox regression and Kaplan-Meier.

### Results

Overall, 72 of 637 patients (11%) had symptomatic VTE; 38 (6%) developed a PE–eight (1.3%) of which were fatal–and 40 (6%) a DVT. After controlling for relevant confounding variables such as age, the modified Charlson Comorbidity Index, visceral metastases, and chemoprophylaxis, longer duration of surgery was independently associated with an increased risk of symptomatic VTE (odds ratio 1.15 for each additional hour of surgery; 95% confidence interval [CI], 1.04-1.28; p=0.009). After controlling for relevant confounding variables such as age, the modified Charlson Comorbidity Index, visceral metastases, and primary tumor type, patients with symptomatic VTE had a worse 1-year survival rate (VTE, 38%; 95% CI, 27–49 versus nonVTE, 47%; 95% CI, 42–51; p=0.044). After controlling for relevant confounding variables, no association was found between wound complications and the use of chemoprophylaxis (odds ratio, 1.34; 95% CI, 0.62–2.90; p=0.459). The overall proportion of patients who developed a wound complication was 10% (66 of 637), including 1.1% (seven of 637) spinal epidural hematomas.

### Conclusion

The risk of both symptomatic PE and fatal PE is high in this patient population, and those with symptomatic VTE were less likely to survive 1-year than those who did not, though this may reflect overall infirmity as much as anything else, because many of these patients did not die from VTE-related complications. Further study, such as randomized controlled trials with consistent postoperative VTE screening comparing different chemoprophylaxis regimens, are needed to identify better VTE prevention strategies.

# INTRODUCTION

Venous thromboembolism (VTE)–encompassing both deep venous thrombosis (DVT) and pulmonary embolism (PE)–is a leading cause of death in patients with cancer.[1–3] Patients with cancer are thought to have a four- to sevenfold increased risk for developing a symptomatic VTE when compared with patients without cancer.[4,5] Spine surgery is also known to be an independent risk factor for symptomatic VTE.6,7 Consequently, patients undergoing surgery for spine metastases may be at an even greater risk for developing these complications.[8] In a recent study among patients undergoing surgery for tumors, 22% developed VTE.[9]

Determining which factors are associated with symptomatic VTE and assessing the consequences of these conditions on survival may identify those who could maximally benefit from symptomatic VTE prevention strategies in the perioperative period. The role of chemoprophylaxis in preventing symptomatic VTE in patients undergoing high-risk spine surgery remains controversial.[8,10] Spine surgeons must weigh the risk of chemoprophylaxis, which includes a higher risk of bleeding, with the benefits of preventing symptomatic VTE. When considering this relationship, one must have an accurate understanding of the incidence and consequences of symptomatic VTE. The incidence of postoperative spinal epidural hematomas and symptomatic VTE–regardless of the use of chemical anticoagulants–appears to be low after spine surgery[8,11], but it is unclear if the same is true after spine metastases surgery.

In this study, we therefore sought to (1) identify the proportion of patients who develop symptomatic VTE within 90-days of surgical treatment for spine metastases; (2) identify the factors associated with the development of symptomatic VTE among patients receiving surgery for spine metastases; (3) assess the association between the development of postoperative symptomatic VTE and 1-year survival among patients who underwent surgery for spine metastases; and (4) assess if chemoprophylaxis increases the risk of wound complications among patients who underwent surgery for spine metastases.

# METHODS

### Study Design and Setting

Our institutional review board approved a waiver of consent between January 2002 and January 2014 for this retrospective study at the authors' hospitals. We included 637 patients who were 18 years of age or older and had surgery for cervical, thoracic, or lumbar metastases (inclusive of lymphoma and multiple myeloma).[12] We excluded patients with (1) kyphoplasty and vertebroplasty; (2) revision procedures, defined as any subsequent procedure after the index surgery addressing the metastatic lesion; and (3) a symptomatic and confirmed VTE within 2 weeks before surgery because this would interfere with the main aim of the study to identify the risk of developing postoperative VTE.

During the period in question, we generally used either 40 mg of enoxaparin or 5000 IUs subcutaneous heparin every 12 hours. Other general thromboembolic prophylaxis approaches and dosages used were: 325 mg of aspirin, 5000 IUs dalteparin daily, and warfarin dependent to maintain an international normalized ratio of 2.0:2.5. There may have been between-surgeon differences in prescribing choices; we found that three surgeons prescribed more heparin relative to LMWH than others. We therefore performed a more-detailed analysis on the patient populations treated by these surgeons and found them not to be different in terms of age, sex, body mass index (BMI), modified Charlson Comorbidity Index, or American Spinal Injury Association (ASIA) impairment scale (more details on this below, in Statistical Analysis, and in Appendix 1).

Patients on preoperative chemoprophylaxis continued their initial medication postoperatively. All chemoprophylaxis was started 48 hours after surgery and continued day to day but was discontinued if a bleeding complication developed. When chemoprophylaxis was contraindicated, an inferior vena cava (IVC) filter was placed before surgery. Chemical anticoagulants, given postoperatively with a maximum range of 14 days, were considered prophylactic. When regimens overlapped, the most aggressive chemoprophylaxis regimen was considered in our analysis.

Mechanical prophylaxis was routinely employed in the form of sequential compression devices and compression stockings at both institutions in all patients during the hospitalization period and therefore not included as a potential treatment variable. Low molecular weight heparin (including enoxaparin and dalteparin, in general dosages of respectively 40 mg and 5000 IUs daily) was the most used chemoprophylaxis in 308 (48%) patients. Subcutaneous heparin was injected into 127 patients (20%); aspirin was used for 92 patients (14%) patients; and warfarin was administered in 21 patients (3.3%). No form of chemoprophylaxis was prescribed for 89 patients (14%).

### Outcomes and Explanatory Variables

Our primary outcome, VTE, was defined as any symptomatic pulmonary embolism (PE) or symptomatic distal or proximal deep venous thromboembolism (DVT) that occurred within 90 days of surgery, presenting with swelling, calf tenderness, tachycardia, pain in the lower extremities, haemoptysis, or tachypnoea. Patients were tested at the earliest possible opportunity after the development of these symptoms. The diagnosis was confirmed by one of the following diagnostic procedures: pulmonary arteriography, vascular ultrasound, venography or chest CT. Our secondary outcome was survival after surgery. The date of death was obtained from the Social Security Index and medical charts. Our third outcome was a documented wound complication within 90 days of surgery, defined as a wound complication that might be attributable to chemoprophylaxis and resulted in longer hospitalization. We excluded wound complications such as wound inflammation resulting in the use of antibiotics. Sixty-six patients (10%) had a documented wound complication, consisting of 34 deep infections (5.3%) that were treated by irrigation and débridement; 28 superficial wound complications (4.4%), such as wound dehiscence resulting in surgery; and 12 deep wound complications (1.9%), including seven symptomatic spinal epidural hematomas (1.1%), four seromas (0.6%), and one splenic bleed 9 days postoperatively (0.2%). Patients with symptomatic spinal epidural hematomas presented with back pain and/or progressive neurologic dysfunction for which the diagnosis was confirmed with MRI. Decompression surgery resulted in six neurologically intact patients (0.9%) and one patient (0.2%) who maintained a deteriorated neurological outcome. Ten patients (1.6%) had a wound complication followed by a symptomatic VTE.

Preoperative local radiotherapy was administered in 30 patients (45%). Disease factors included primary tumor type, pathologic fracture, number of bone and visceral metastases, preoperative ASIA impairment scale, and time from primary tumor diagnosis to operation for metastatic disease. Clinical factors included the preoperative comorbidity status defined using the modified Charlson Comorbidity Index[13] and the use of preoperative local radiotherapy or systemic therapy. Treatment factors included prior embolization, IVC filter placement, vertebral levels included in surgery, surgical approach, surgery duration in hours, estimated blood loss during surgery (liters), total perioperative transfusions, and hospitalization days. Laboratory factors included preoperative hemoglobin levels (g/dL), preoperative white blood cell count (1000/mm3), preoperative platelet count (1000/mm3), creatinine levels (mg/dL) and calcium levels (mg/dL). We obtained preoperative laboratory values by choosing laboratory values nearest to surgery with a maximum range of 7 days. We used the modified Charlson Comorbidity Index to determine the comorbidity status based on an algorithm of the ICD-9 codes classifying 12 comorbidities. We dichotomized the comorbidity status into any additional comorbidity or none (in addition to the metastases). We determined any preoperative neurological deficits (grade A, B, C, or D) or none (score E, including patients with prior but no present deficits) using the preoperative ASIA impairment scale.[14] The Eastern Cooperative Oncology Group (ECOG) performance status was dichotomized into good (0, 1, or 2) or poor scores (3 to 4).[12,15] We defined previous systemic therapy as all types of non-radiotherapeutic adjuvants or nonsurgical adjuvants, for example, immunologic, cytotoxic, metabolic, or hormonal therapy, administered before surgery. We considered the presence of an IVC filter before surgery or within 90 days postoperatively as prophylactic, except when it was placed after a symptomatic VTE event.

### Demographics, Description of Study Population

The patients included 371 males (58%) and 266 females (42%) with a median age of 60 years (interquartile range [IQR], 52–68 years; Table 1). The median duration of surgery was 6 hours (IQR, 5–8 hours) and median hospitalization was 8 days (IQR, 6–12 days). Of the 637 spine metastatic operations, 371 (58%) involved the thoracic region; 141 (22%) involved the lumbar area; 86 (14%) the cervical; and 39 (6%) the combined region. IVC filters were placed in 41 patients: 34 (6%) in the nonVTE and seven (10%) in the VTE group. Most common primary tumor types included lung (18%), kidney (13%), and breast cancer (12%; Table 2).

**Table 1.** Patient- and treatment characteristics for the no VTE and VTE group (n=637)

| Variables | No VTE (n=565) | VTE (n=72) |
|---|---|---|
| | Median (IQR) | |
| Age (years) | 60 (52-68) | 61 (54 - 68) |
| Modified Charlson Comorbidity Index | 6 (6 - 8) | 6 (6 - 8) |
| Total estimated blood loss during surgery (liters)[a] | 700 (350 - 1400) | 1000 (500 - 1800) |
| Duration surgery (hours)[a] | 369 (271 - 484) | 448 (333 - 567) |
| Duration primary diagnosis till metastatic operation (days) | 398 (25 - 1436) | 337 (4 - 1168) |
| Duration hospitalization (days) | 8 (6 - 12) | 8 (6 - 13) |
| Total perioperative transfused[b] | 2 (0 - 4) | 2 (0 - 5) |
| Preoperative laboratory values[a] | | |
| Hemoglobin levels (g/dL) | 11 (10 - 13) | 11 (10 - 13) |
| White blood cell count ($10^3$/ L) | 11 (7 - 14) | 12 (8 - 16) |
| Creatinine levels (mg/dL) | 0.8 (0.6 - 0.9) | 0.8 (0.6 - 0.9) |
| Calcium levels (mg/dL) | 8.7 (8.0 - 9.2) | 8.7 (7.9 - 9.3) |
| Platelet count ($10^3$/$mm^3$) | 238 (179 - 322) | 249 (176 - 340) |
| | % (n) | |
| Men | 331 (59) | 40 (56) |
| Additional comorbidities[c] | 259 (46) | 34 (47) |
| Body mass index (in kg/$m^2$)[a] | | |
| < 18.5 | 16 (3) | 3 (4) |
| 18.5 – 30 | 354 (73) | 40 (56) |
| > 30 | 113 (23) | 22 (31) |
| Smoking status[a] | | |
| Never smoked | 204 (37) | 26 (36) |
| Former smoker | 244 (45) | 33 (46) |
| Current smoker | 97 (18) | 12 (17) |
| ASIA impairment scale (preoperative) | | |
| Neurological deficit (A, B, C, or D) | 264 (47) | 37 (51) |
| No neurological deficit (E) | 301 (53) | 35 (49) |
| ECOG performance status[a] | | |
| Score 0 to 2 (≤50% of waking hours bed or chair bound) | 308 (80) | 32 (44) |
| Score 3 to 4 (>50% of waking hours bed or chair bound) | 78 (20) | 13 (18) |
| Time between start of neurological symptoms and surgery | | |
| No neurological symptoms | 299 (53) | 37 (51) |
| <14 days | 153 (27) | 22 (31) |
| ≥14 days | 113 (20) | 13 (18) |
| Number of spine levels undergoing operation | | |
| 1 | 357 (63) | 43 (60) |
| 2 | 100 (18) | 12 (17) |
| 3 or more | 108 (19) | 17 (24) |
| Number of spine metastases | | |
| 1 | 152 (27) | 23 (32) |
| 2 | 83 (15) | 9 (13) |
| 3 or more | 330 (58) | 40 (56) |

*Continued on next page*

| | | |
|---|---|---|
| Metastases region | | |
| Thoracic | 324 (57) | 47 (65) |
| Lumbar | 126 (22) | 15 (21) |
| Cervical | 82 (15) | 4 (6) |
| Combined | 33 (6) | 6 (8) |
| Multiple bone metastases | 302 (53) | 36 (50) |
| Visceral metastases | 184 (33) | 26 (36) |
| Prior embolization | 117 (21) | 21 (29) |
| IVC filter prophylaxis | 34 (6) | 7 (10) |
| Previous local radiotherapy | 189 (33) | 29 (40) |
| Previous systemic therapy | 325 (58) | 37 (51) |
| Type of surgery | | |
| Vertebrectomy or corpectomy with stabilization | 268 (47) | 40 (56) |
| Decompression and stabilization | 202 (36) | 25 (35) |
| Decompression | 77 (14) | 6 (8) |
| Stabilization | 18 (3) | 1 (1) |
| Surgical approach | | |
| Posterior | 482 (85) | 61 (85) |
| Anterior | 59 (10) | 7 (10) |
| Combined | 24 (4) | 4 (6) |

*VTE=venous thromboembolism; IQR=interquartile range; ASIA=American Spinal Injury Association; ECOG=Eastern Cooperative Oncology Group; IVC=inferior vena cava.*
*a: Available data in the following variables: estimated blood loss in 499 patients (88%) from the no VTE group and in 67 patients (93%) from the VTE group, duration of surgery in 501 patients (89%) from the no VTE group and in 65 patients (90%) from the VTE group, preoperative hemoglobin level in 555 patients (98%) from the no VTE group, preoperative white blood cell count in 554 patients (98%) from the no VTE group, preoperative creatinine levels in 546 patients (97%) from the no VTE group and in 70 patients (97%) from the VTE group, preoperative calcium level in 513 patients (91%) from the no VTE group and in 70 patients (97%) from the VTE group, preoperative platelet count in 553 patients (98%) from the no VTE group, body mass index in 483 patients (85%) from the no VTE group and in 65 patients (90%) from the VTE group, smoking status in 545 patients (96%) from the no VTE group and in 71 patients (99%) from the VTE group, and ECOG in 386 patients (68%) from the no VTE group and in 45 patients (63%) from the VTE group.*
*b: Total perioperative transfused includes all blood and non-blood products.*
*c: These values were based on any additional comorbidity on top of the metastatic disease score according to the modified Charlson Comorbidity Index.*

### Accounting for All Patients/Study Subjects

We identified 1330 patients using the ICD-9 code for the diagnosis of pathologic vertebrae fracture (733.13). Additionally, we further identified 796 patients using a word-based inquiry of operative reports in our medical database. We manually inspected the medical charts of these 2126 patients for eligibility by two independent research fellows (OQG, PTO), and ultimately included 637 patients.[16] We verified followup until October 4, 2016. At followup after 90 days and 1 year, respectively, 21 of 637 (3%) and 41 of 637 (6%) were lost to followup.

### Statistical Analysis

We used multivariable logistic regression analysis controlling for confounding variables with a p value < 0.10 from bivariate testing and presumed to be relevant to VTE to assess independent risk factors for symptomatic VTE.[9] Odds ratios for continuous variables are interpreted in terms of each

**Table 2.** Origin of primary tumor (n=637).

| *Primary tumor* | *% (n)* |
|---|---|
| Lung | 113 (18) |
| Kidney | 80 (13) |
| Breast | 76 (12) |
| Multiple myeloma | 71 (11) |
| Prostate | 56 (9) |
| Melanoma | 27 (4) |
| Colorectal | 23 (4) |
| Neuroendocrine | 21 (3) |
| Sarcoma | 21 (3) |
| Lymphoma | 19 (3) |
| Head and neck | 17 (3) |
| Thyroid | 15 (2) |
| Hepatocellular | 13 (2) |
| Esophageal | 11 (2) |
| Endometrial | 11 (2) |
| Salivary gland | 6 (1) |
| Adenocarcinoma | 5 (1) |
| Bladder | 5 (1) |
| Other* | 52 (8) |

**This category included unknown cancer 20 (3.1%), pancreatic cancer 4 (0.6%), germ cell cancer 4 (0.6%), ovarian cancer 3 0.5%), testicular cancer 3 (0.5%), penile cancer 3 (0.5%), cholangiocarcinoma 3 (0.5%), gastric cancer 2 (0.3%), adrenal cancer 2 (0.3%), blue-cell tumor 1 (0.2%), skin cancer 1 (0.2%) and leukemia 1 (0.2%).*

unit increase or decrease on the scale (that is, 1 to 2, 2 to 3, etc; each one-unit/hour increment of longer duration of surgery with an odds ratio of > 1 corresponds to an increased risk of the outcome in question, in this case, symptomatic VTE). In bivariate analysis, two of the 33 variables examined were associated with increased and decreased risk of symptomatic VTE development (Appendix 2), respectively: longer duration of surgery (odds ratio [OR], 1.17; 95% confidence interval (CI), 1.07–1.28; p=0.001) and the absence of metastatic lesion in the cervical region (OR, 0.34; 95% CI, 0.12–0.96; p=0.042). Two additional variables with P-values < 0.10 were included in the multivariable analysis: preoperative white blood cell count (OR, 1.03; 95% CI, 1.00–1.07; p=0.064) and BMI of > 30 kg/m2 (OR, 1.72; 95% CI, 0.98–3.02; p=0.058). Additional variables controlled for were age, the modified Charlson Comorbidity Index, the preoperative ASIA impairment scale, visceral metastases, chemoprophylaxis, and estimated blood loss during surgery. Multivariate logistic regression was also performed to examine the association between chemoprophylaxis and wound complication controlling for age, sex, and the modified Charlson Comorbidity Index. Bivariate logistic regression was used to assess chemoprophylaxis regimens and surgeons (Appendix 1).

The following baseline characteristics were assessed for differences between surgeons, none were significant: age (p=0.821, sex (p=0.344), BMI (p=0.067), the modified Charlson Comorbidity Index (p=0.077, and ASIA impairment scale (p=0.253). We used Cox regression analysis after controlling for the confounding factors age, sex, BMI, the modified Charlson Comorbidity Index, visceral metastases, primary tumor type, previous systemic therapy, and operation type to assess a difference in survival between the symptomatic VTE and nonVTE group. Kaplan-Meier plots demonstrated the survival curves for both groups. Multiple chained imputation was used to estimate missing values to retain all values for multivariable analysis. The dataset was recreated multiples times (40 in our cohort) by multiple imputation and the missing values were estimated with plausible values based on the residual variables accounting for uncertainty. Statistical software estimated missing values for: BMI (14% [89 of 637]), duration of surgery (11% [71 of 637]), total estimated blood loss during surgery (11% [71 of 637]), and preoperative white blood cell count (1.7% [11 of 637]). Two-tailed p values of < 0.05 were considered significant. We used Stata 13 (StataCorp LP, College Station, TX, USA) to perform all statistical analyses.

# RESULTS

Symptomatic VTE was diagnosed in 72 patients (11%), DVT in 40 patients (6.2%), and PE in 38 patients (6.0%) within 90 days after spine surgery for metastases (Table 3). The median age of the 72 patients was 61 years (IQR, 54–68 years), and 40 patients (56%) were men. Six patients (0.9%) had concurrent evidence for a PE and DVT, and eight (1.3%) PEs were fatal (Table 4). Most symptomatic VTEs developed after hospital discharge: the median time between surgery and symptomatic VTE was 21 days (IQR, 7–39 days), and the median postoperative hospitalization of these patients was 8 days (IQR, 6–13 days). The median postoperative day of developing a fatal PE was 24 (IQR, 11–41 days).

After controlling for potentially relevant confounding variables such as age, the modified Charlson Comorbidity Index, visceral metastases, and chemoprophylaxis, we found that longer duration of surgery (OR, 1.15 for each additional hour of surgery; 95% CI, 1.04–1.28; p=0.009) was independently associated with the development of symptomatic VTE (Table 5). Symptomatic VTE developed in 42 of 374 patients (11%) in the group that used any chemoprophylaxis and in 30 of 263 patients (11%) who received no chemoprophylaxis, demonstrating no association after controlling for age, sex, and the Modified Charlson Comorbidity Index (OR, 0.96; 95% CI, 0.58–1.59; p=0.863).

Patients with symptomatic VTE compared with those without had lower 1-year survival after controlling for the following potentially confounding variables: age, sex, BMI, the modified Charlson Comorbidity Index, visceral metastases, primary tumor type, previous systemic therapy, and operation type (VTE: 38%; 95% CI, 27–49 versus without VTE: 47%; 95% CI, 42–51; p=0.044; Figure 1). The probability of developing a symptomatic VTE rose gradually over the 90-day postoperative period with a notable increase at 30 days after surgery (Figure 2). Timing of fatal PEs ranged from postoperative day 1 to 78.

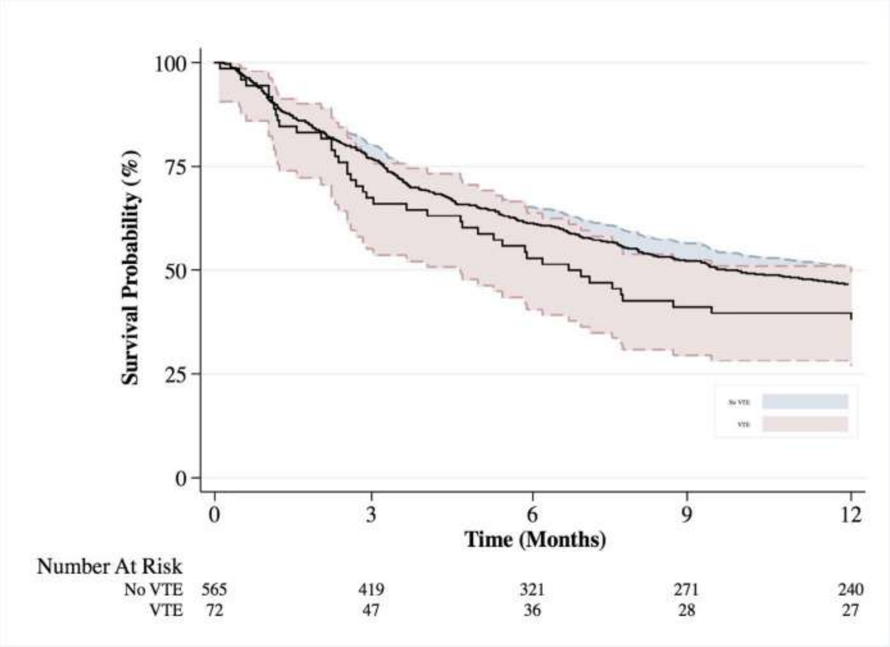


**Figure 1.** This Kaplan-Meier plot shows the survival probability with 95% CIs for patients with and without postoperative symptomatic VTE (p=0.044).

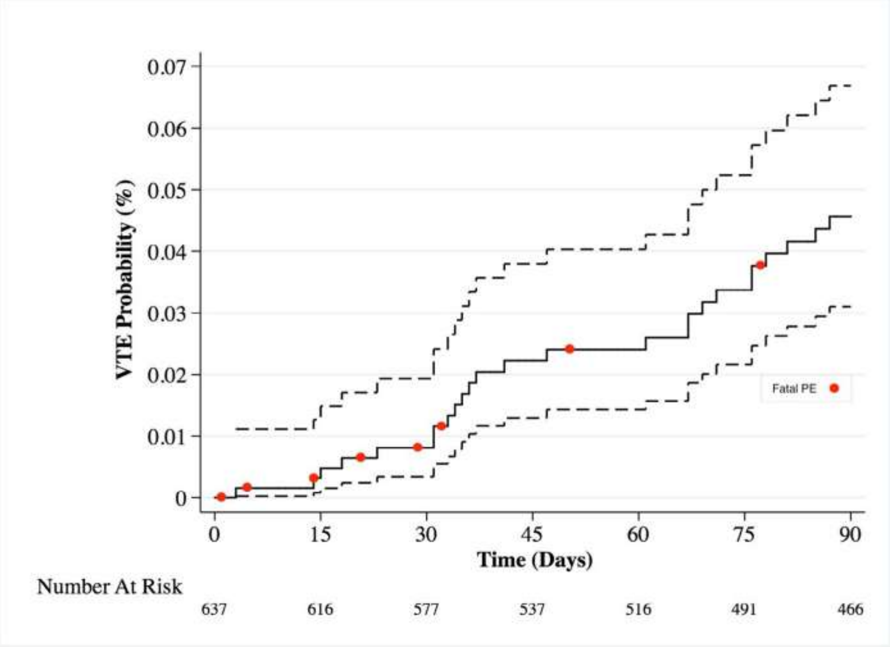


**Figure 2.** This Kaplan-Meier plot demonstrates the probability of developing a symptomatic VTE within 90 days postoperatively 0.05% (95% CI, 0.03–0.07). The risk of developing symptomatic VTE had a sudden increase at 30 days postoperatively and kept rising gradually thereafter.

After controlling for age, sex, and the modified Charlson Comorbidity Index, we found no association between any of the different chemoprophylaxis regimens and the occurrence of 66 wound complications, consisting of 31 (47%) for LMWH (reference value), 17 (26%) for subcutaneous heparin (OR, 0.94; 95% CI, 0.21–4.23; p=0.936), nine (14%) for aspirin (OR, 0.97; 95% CI, 0.44–2.1 p=0.937), seven (11%) for no form of chemoprophylaxis (OR, 0.76; 95% CI, 0.32–1.80; p=0.535), and two (3%)

**Table 3.** 90-Days symptomatic VTE, wound complications and anticoagulant (n=637)

| *Variables* | *All patients (n=637)* |
|---|---|
| | *n (%)* |
| VTE events[a] | 72 (11) |
| DVT | 40 (6.2) |
| PE | 38 (6.0) |
| Wound complications[a] | 66 (10) |
| Deep infection | 34 (5.3) |
| Superficial wound complication | 28 (4.4) |
| Deep wound complication | 12 (1.9) |
| Anticoagulant | |
| None | 89 (14) |
| Low molecular weight heparin | 308 (48) |
| Subcutaneous heparin | 127 (20) |
| Aspirin | 92 (14) |
| Warfarin | 21 (3.3) |
| Wound complication followed by VTE[a] | 10 (1.6) |
| Superficial infection | 0 (0) |
| Deep infection | 4 (0.6) |
| Superficial wound complication | 6 (0.9) |
| Deep wound complication | 2 (0.3) |
| | *Median (IQR)* |
| Time between surgery and VTE (days) | 21 (7 - 39) |
| DVT | 21 (8 - 42) |
| PE | 22 (7 - 43) |
| PE fatal | 24 (11 - 41) |
| Duration postoperative hospitalization for VTE (days) [b] | 8 (6 - 13) |
| DVT | 8 (6 - 12) |
| PE | 8 (6 - 14) |
| PE fatal | 14 (9 -20) |
| Time between surgery and wound complication (days) | 21 (9 -34) |
| Superficial infection | 40 (5 - 44) |
| Deep infection | 25 (19 - 34) |
| Superficial wound complication | 31 (18 - 36) |
| Deep wound complication | 8 (3 - 14) |

*VTE=venous thromboembolism; DVT=deep vein thrombosis; PE=pulmonary embolism; IQR=interquartile range.*
*a Not mutually exclusive.*
*b The duration of hospitalization is the time in days between surgery and discharge, not specifically the hospitalization for VTE management.*

for warfarin (OR, 1.38; 95% CI, 0.73–2.60; p=0.316). Likewise, an additional sub analysis between the usage of a chemical anticoagulant versus no chemical anticoagulant showed no difference, but this was underpowered.

**Table 4.** Multivariate logistic regression assessing risk factors for 90-day symptomatic VTE after multiple imputation (40 imputations) (n=637)

| *Variables* | *Odds ratio (95% CI)* | *Standard error* | *P-value* |
|---|---|---|---|
| Age | 1.00 (0.97 - 1.02) | 0.01 | 0.600 |
| Modified Charlson Comorbidity Index | 1.09 (0.93-1.29) | 0.09 | 0.278 |
| Body mass index (in kg/m$^2$)[a] | | | |
| < 18.5 | 1.94 (0.50 - 7.57) | 1.35 | 0.336 |
| 18.5 – 30 | | *Reference value* | |
| > 30 | 1.46 (0.80-2.67) | 0.45 | 0.218 |
| ASIA impairment scale (preoperative) | | | |
| Neurological deficit (A, B, C, or D) | 1.37 (0.80-2.35) | 0.38 | 0.253 |
| No neurological deficit (E) | | *Reference value* | |
| Metastases region | | | |
| Thoracic | | *Reference value* | |
| Lumbar | 0.98 (0.50-1.91) | 0.33 | 0.956 |
| Cervical | 0.40 (0.13-1.20) | 0.22 | 0.103 |
| Combined | 1.17 (0.44-3.08) | 0.58 | 0.750 |
| Preoperative white blood cell count ($10^3$/ L)[a] | 1.03 (0.99-1.07) | 0.02 | 0.198 |
| Visceral metastases | 1.09 (0.61-1.92) | 0.32 | 0.776 |
| Duration of surgery (hours)[a] | 1.15 (1.04-1.28) | 0.06 | **0.009** |
| Total estimated blood loss during surgery (liters)[a] | 0.92 (0.75-1.13) | 0.10 | 0.420 |
| Anticoagulant | | | |
| None | 1.36 (0.57-3.26) | 0.61 | 0.487 |
| Low-molecular-weight heparin | | *Reference value* | |
| Subcutaneous heparin | 1.97 (0.57-6.73) | 1.23 | 0.282 |
| Aspirin | 1.43 (0.66-3.11) | 0.57 | 0.368 |
| Warfarin | 1.17 (0.57-2.41) | 0.43 | 0.662 |

*BMI=body mass index; VTE=venous thromboembolism; CI=confidence interval; ASIA=American Spinal Injury Association.*
**Bold** *indicates significance (P<0.05)*
*a BMI in 483 patients (85%) from the no VTE group and in 65 patients (90%) from the VTE group, preoperative white blood cell count in 554 patients (98%) from the no VTE group and in 72 patients (100%) in the VTE group, duration of surgery in 501 patients (89%) from the no VTE group and in 65 patients (90%) in the VTE group, and estimated blood loss was available in 499 patients (88%) from the no VTE group and in 67 patients (93%) from the VTE group.*

Table 5. Patient characteristics for fatal PE events

| Sex, age (years) | BMI (kg/m$^2$) | Primary tumor histology | Location of metastases | Radiation | Chemo-therapy | Anticoagulant | Postoperative day | Hospitalization (days)[a] | Wound complication | Type of surgery |
|---|---|---|---|---|---|---|---|---|---|---|
| F, 65 | 23 | Lung | Thoracic | No | No | Aspirin | 27 | 5 | None | Decompression |
| F, 57 | - | Lung | Combined | No | No | Low-molecular-weight heparin | 78 | 14 | None | Decompression and stabilization |
| M, 64 | - | Kidney | Thoracic | No | No | Warfarin | 31 | 32 | None | Vertebrectomy or corpectomy with stabilization |
| F, 26 | - | Multiple Myeloma | Thoracic | No | Yes | None | 1 | 13 | None | Decompression and stabilization |
| F, 47 | 25 | Kidney | Combined | No | Yes | Heparin | 15 | 18 | Deep wound complication | Vertebrectomy or corpectomy with stabilization |
| F, 57 | 31 | Germ cell | Lumbar | Yes | Yes | None | 6 | 22 | None | Vertebrectomy or corpectomy with stabilization |
| F, 50 | 41 | Breast | Thoracic | Yes | Yes | Low molecular weight heparin | 50 | 7 | None | Decompression and stabilization |
| M, 54 | 35 | Lung | Thoracic | No | No | None | 21 | 10 | None | Vertebrectomy or corpectomy with stabilization |

*PE=pulmonary embolism; VTE=venous thromboembolism; BMI=body mass index; F=female; M=male.*
*a The duration of hospitalization is the time in days between surgery and discharge, not specifically the hospitalization for VTE management.*

# DISCUSSION

Malignant disease and surgery are two major risk factors for symptomatic VTE[17-19], and the development of symptomatic VTE in patients with cancer is associated with poor survival.[20] Spine surgeons must weigh the risk of chemoprophylaxis, which includes hemorrhagic complications, with the benefits of preventing symptomatic VTE in patients undergoing surgery for spine metastases. Our goal in this study was to investigate the risk of symptomatic VTE, the association between postoperative symptomatic VTE development and 1-year survival and assess the relationship between chemoprophylaxis and proportion of wound complications. A total of 11% of patients developed symptomatic VTE (including 6% who developed symptomatic PE); 1.3% of the patients in this series died of PE. After controlling for potential confounding variables, we found that longer duration of surgery was independently associated with an increased risk of symptomatic VTE and that patients with symptomatic VTE had worse 1-year survival. We did not find an association between the usage of chemical anticoagulants and the development of postoperative wound complications or symptomatic VTE. However, our study was underpowered to show this difference.

This study had several limitations. The most important limitation in this series was the inconsistent use of chemoprophylaxis regimens. In general, we used either 40 mg of enoxaparin or 5000 IUs subcutaneous heparin every 12 hours. Other general thromboembolic prophylactic dosages used were: 325 mg of aspirin, 5000 IUs dalteparin daily, and warfarin dependent to maintain an international normalized ratio of 2.0:2.5. Additional analysis demonstrated that three surgeons prescribed more heparin relative to LMWH than others, despite the lack of identifiable differences in baseline characteristics (Appendix 1); this was most likely based on personal preference of these specific surgeons. While an obvious limitation, LMWH likely was chosen over heparin due to more predictable pharmacokinetics and fewer nonhemorrhagic side-effects.[21] A meta-analysis in medically ill patients showed no difference in major bleeding events between the two, which leads us to believe that the different prescribing pattern of these surgeons likely had only a negligible impact on our study's results.[22]

Second, screening for and detection of VTE may have been inconsistent over the years, where the threshold for screening may have been lower in more recent years and detection techniques more effective. However, year of surgery was not associated with VTE occurrence, and no differences were found in baseline characteristics of the patient population over the years. Also, the proportion of patients with VTE may have been underestimated given the lack of a universal screening protocol and the fact that we were only able to include those with symptomatic events. We anticipate a relatively low number of missed events because these patients are part of a complicated group that remains under enhanced postoperative surveillance. Third, the survival analysis between the symptomatic VTE and nonVTE groups likely consists of uncontrolled for differences. However, we controlled for the most important confounding survival variables, such as age, primary tumor type, and the modified Charlson Comorbidity Index. Fourth, the exact duration and compliance of anticoagulation could not always be confirmed. However, both centers have implemented protocols that call for patients to continue anticoagulation regimens 4 weeks after surgery and employ sequential compression devices and compression stockings during hospitalization. Fifth, history of VTE was excluded in the analysis because of the unreliability of this specific personal history data in a tertiary center. Sixth, lymphoma and multiple myeloma metastasized to the spine were included, which are known for their better prognosis and this could have potentially led to selection bias.[23] Nonetheless, these patients represent a sizeable portion of patients, 90 of 637 (14%), who develop spine metastasis and therefore warrant consideration in a study such as this. Seventh, this cohort represents a heterogeneous population with numerous comorbidities and potential confounders. We attempted to control for these factors by using multivariable regression testing and the Charlson Comorbidity Index, ECOG performance status, and ASIA impairment scale as objectification of case complexity, yet we recognize the prospect of residual confounding. Lastly, this remains a retrospective work with all the inherent limitations associated with such a study design, including reliance on chart abstraction and search algorithms to identify eligible patients.

Compared with previous work, this study has a relatively large sample of patients with spinal metastases collected over the last 15 years and extensive followup considering the cohort's clinical characteristics.[9,24] In this series of 637 patients, 11% developed symptomatic VTE (72 of 637) with 6.0% (38 of 637) developing PE, and 1.3% (eight of 637) who died of this complication. Another, similar study reported a symptomatic VTE in 1.6% of patients undergoing surgery for symptomatic spinal metastases, but this study was designed to determine survival and not specifically evaluate the risk of symptomatic VTE [24]. Compared with similar VTE studies for spine and musculoskeletal tumor surgery, the findings are relatively high; particularly, the number of fatal PEs is unprecedented.[10,18,25–33] Multiple factors might explain the high risk of symptomatic VTE observed here, including older age, neoplasm, severe venous stasis, prolonged immobilization and paralysis postoperatively, as well as longer operation times.[7,10,34]

Longer duration of spine surgery was independently associated with an increased risk of postoperative symptomatic VTE in our series, a surgical factor postulated in previous surgical research.[35] Clinically, procedures with prolonged surgery, where each hour corresponds with an increased odds ratio of 1.15 for symptomatic VTE development, may warrant greater consideration for symptomatic VTE prevention such as chemoprophylaxis. Some tumor histologies, especially lymphoma and multiple myeloma, are proven to be associated with hypercoagulability and increased symptomatic VTE risk[36], but none of them were identified as factors associated with VTE. However, we were not sufficiently powered to address this association.

The development of postoperative VTE has an association with a decreased 1-year survival rate. This poor survival in patients with symptomatic VTE can be explained by the highly complex patient population with multiple comorbidities and other disease-related factors, in which patients with more advanced cancer develop VTE more easily. VTE may function more as an infirmity marker than as the main cause for poor survival. Another explanation may be the high incidence of fatal PEs (1.3%), suggesting that VTE prevention, such as adequate chemoprophylaxis, could improve short-term survival. However, spine surgeons may be hesitant to use chemical anticoagulants after spine surgery out of concern for severe hemorrhagic complications, such as spinal epidural hematoma.[11,16] A recent study reported that aspirin is safe regarding wound complications and effective in preventing symptomatic VTE after total joint arthroplasty, which makes it a viable chemoprophylaxis agent in the spine metastases population.[37] Therefore, given the incidence of fatal PEs (1.3%) and symptomatic spinal epidural hematomas (1.1%; including only one (0.2%) patient who developed a deteriorated neurological outcome), further study is desirable to assess more adequate anticoagulation in this population.

With respect to the timing of symptomatic VTE events, risk increased after 30 days postoperatively and kept rising (Figure 2). In addition, half the symptomatic VTEs occurred after 3 weeks (21 days; IQR, 7–39 days), which is considerably longer compared with postoperative hospitalization for VTE patients (8 days; IQR 6–13 days). We also observed that symptomatic VTE timing exceeded hospitalization in fatal PEs (respectively, 24; IQR, 11–41 versus 14; IQR, 9–20). Similar studies have reported comparable results about this late onset of symptomatic VTE.[25–27,29,32,33] Both institutions followed protocols that recommend postoperative anticoagulant regimens of about 4 weeks, but considering the late onset of symptomatic VTE, a longer duration of anticoagulant use may be indicated.[27,33,38] The national orthopaedic guidelines are unclear about addressing this problem of chemoprophylaxis duration stating that the “patients and physicians discuss the duration of prophylaxis”.[39] Spine surgeons have demonstrated practice variability in high-risk spine surgery patients regarding not only the duration of chemoprophylaxis, but also the use of chemoprophylactic agents.[8] In addition, it is also possible that there is a gap between actual received outpatient chemoprophylaxis and guideline recommendations. Although this study was not designed to specifically address this compliance variable, previous studies report poor compliance in outpatient anticoagulant prophylaxis after major orthopaedic surgery and prophylaxis prescription at discharge.[40–42] A clear trend is developing toward shorter hospitalizations after major orthopaedic surgery[43], necessitating more emphasis on compliance of outpatient anticoagulant prophylaxis. Novel oral anticoagulants may fulfil a prominent role, since most patients prefer oral agents.[44] Further study, preferably a randomized control trial with consistent postoperative VTE screening, could help surgeons better balance the risks and benefits as they choose from among the available postoperative prophylactic regimens.

# CONCLUSION

This study demonstrates a high risk of symptomatic 90-day VTE among patients undergoing spine surgery for metastases; 11% of the patients developed symptomatic VTE, 6% developed a symptomatic pulmonary embolism, and 1.3% patients died of that complication. While those with symptomatic VTE were less likely to survive 1-year than those who did not, we recognize that this may reflect overall infirmity as much as anything else, since many of these patients did not die from complications related to VTE. Further studies such as randomized controlled trials with consistent postoperative VTE screening comparing different chemoprophylaxis regimens are required to identify better symptomatic VTE prevention.

# REFERENCES

1. Ambrus JL, Ambrus CM, Mink IB, et al. Causes of death in cancer patients. *J Med.* 1975;6(1):61–4.

2. Donati MB. Cancer and thrombosis. *Haemostasis.* 1994;24(2):128–31.

3. Khorana AA, Francis CW, Culakova E, et al. Thromboembolism is a leading cause of death in cancer patients receiving outpatient chemotherapy. *J Thromb Haemost.* 2007;5(3):632–634.

4. Blom JW. Malignancies, prothrombotic mutations, and the risk of venous thrombosis. *JAMA.* 2005;293(6):715.

5. Heit JA, Silverstein MD, Mohr DN, et al. Risk factors for deep vein thrombosis and pulmonary embolism: a population-based case-control study. *Arch Intern Med.* 2000;160(6):809–15.

6. Clagett GP, Anderson Jr. FA, Geerts W, et al. Prevention of venous thromboembolism. *Chest.* 1998;114(5 Suppl):531S-560S.

7. Geerts WH, Bergqvist D, Pineo GF, et al. Prevention of venous thromboembolism: American College of Chest Physicians evidence-based clinical practice guidelines (8th edition). *Chest.* 2008;133(6.6):381S-453S.

8. Glotzbecker MP, Bono CM, Harris MB, et al. Surgeon practices regarding postoperative thromboembolic prophylaxis after high-risk spinal surgery. *Spine (Phila. Pa. 1976).* 2008;33(26):2915–2921.

9. Yoshioka K, Murakami H, Demura S, et al. Comparative Study of the Prevalence of Venous Thromboembolism After Elective Spinal Surgery. *Orthopedics.* 2013;36(2):e223–e228.

10. Glotzbecker MP, Bono CM, Wood KB, et al. Thromboembolic disease in spinal surgery: a systematic review. *Spine (Phila. Pa. 1976).* 2009;34(3):291–303.

11. Glotzbecker MP, Bono CM, Wood KB, et al. Postoperative spinal epidural hematoma: a systematic review. *Spine (Phila Pa 1976).* 2010;35(10):E413-20.

12. Nathan SS, Healey JH, Mellano D, et al. Survival in patients operated on for pathologic fracture: Implications for end-of-life orthopedic care. *J Clin Oncol.* 2005;23(25):6072–6082.

13. Quan H, Li B, Couris CM, et al. Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol.* 2011;173(6):676–682.

14. Kirshblum SC, Burns SP, Biering-Sorensen F, et al. International standards for neurological classification of spinal cord injury (Revised 2011). *J Spinal Cord Med.* 2011;34(6):535–546.

15. Oken M, Creech R, Tormey D, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol.* 1982;5(6):649–656.

16. Paulino Pereira NR, Ogink PT, Groot OQ, et al. Complications and reoperations after surgery for 647 patients with spine metastatic disease. *Spine J* 2019;19(1):144–156.

17. Baron JA, Gridley G, Weiderpass E, et al. Venous thromboembolism and cancer. *Lancet (London, England).* 1998;351(9109):1077–80.

18. Benevenia J, Bibbo C, Patel D V, et al. Inferior vena cava filters prevent pulmonary emboli in patients with metastatic pathologic fractures of the lower extremity. *Clin Orthop Relat Res.* 2004;(426):87–91.

19. Lee AYY. Cancer and venous thromboembolism: prevention, treatment and survival. *Journal of Thrombosis and Thrombolysis*.Vol 25.; 2008:33–36.

20. Sørensen HT, Mellemkjær L, Olsen JH, et al. Prognosis of cancers associated with venous thromboembolism.

*N Engl J Med.* 2000;343(25):1846–1850.

21. Garcia DA, Baglin TP, Weitz JI, et al. Parenteral Anticoagulants. *Chest.* 2012;141(2):e24S-e43S.

22. Kanaan AO, Silva MA, Donovan JL, et al. Meta-analysis of venous thromboembolism prophylaxis in medically Ill patients. *Clin Ther.* 2007;29(11):2395–2405.

23. Katagiri H, Okada R, Takagi T, et al. New prognostic factors and scoring system for patients with skeletal metastasis. *Cancer Med.* 2014;3(5):1359–1367.

24. Amelot A, Balabaud L, Choi D, et al. Surgery for metastatic spine tumors in the elderly. Advanced age is not a contraindication to surgery! *Spine J* 2017;17(6):759–767.

25. Tuy B, Bhate C, Beebe K, et al. IVC filters may prevent fatal pulmonary embolism in musculoskeletal tumor surgery. *Clin Orthop Relat Res.* 2009;467(1):239–245.

26. Damron TA, Wardak Z, Glodny B, et al. Risk of venous thromboembolism in bone and soft-tissue sarcoma patients undergoing surgical intervention: A report from prior to the initiation of SCIP measures. *J Surg Oncol.* 2011;103(7):643–647.

27. Groot OQ, Ogink PT, Janssen SJ, et al. High Risk of Venous Thromboembolism After Surgery for Long Bone Metastases. *Clin Orthop Relat Res.* 2018;476(10):2052–2061.

28. Lin PP, Graham D, Hann LE, et al. Deep venous thrombosis after orthopedic surgery in adult cancer patients. *J Surg Oncol.* 1998;68(1):41–7.

29. Mitchell SY. Venous thromboembolism in patients with primary bone or soft-tissue sarcomas. *J Bone Jt Surg.* 2007;89(11):2433.

30. Morii T, Mochizuki K, Tajima T, et al. Venous thromboembolism in the management of patients with musculoskeletal tumor. *J Orthop Sci.* 2010;15(6):810–815.

31. Nathan SS, Simmons KA, Lin PP, et al. Proximal deep vein thrombosis after hip replacement for oncologic indications. *J Bone Joint Surg Am.* 2006;88(5):1066–70.

32. Patel AR, Crist MK, Nemitz J, et al. Aspirin and compression devices versus low-molecular-weight heparin and PCD for VTE prophylaxis in orthopedic oncology patients. *J Surg Oncol.* 2010;102(3):276–281.

33. Shallop B, Starks A, Greenbaum S, et al. Thromboembolism after intramedullary nailing for metastatic bone lesions. *J Bone Jt Surgery-American Vol.* 2015;97(18):1503–1511.

34. Al-Dujaili TM, Majer CN, Madhoun TE, et al. Deep venous thrombosis in spine surgery patients: Incidence and hematoma formation. *Int Surg.* 2012;97(2):150–154.

35. Kim JYS, Khavanin N, Rambachan A, et al. Surgical duration and risk of venous thromboembolism. *JAMA Surg.* 2015;150(2):110.

36. DeStefano V, Za T, Rossi E. Venous thromboembolism in multiple myeloma. *Semin Thromb Hemost.* 2014;40(3):338–347.

37. Parvizi J, Huang R, Restrepo C, et al. Low-dose aspirin is effective chemoprophylaxis against clinically important venous thromboembolism following total joint arthroplasty: A preliminary analysis. *J Bone Joint Surg Am.* 2017;99(2):91–98.

38. Sweetland S, Green J, Liu B, et al. Duration and magnitude of the postoperative risk of venous thromboembolism in middle aged women: prospective cohort study. *BMJ.* 2009;339(dec03 1):b4583–b4583.

39. Mont MA Jacobs JJ et al. **Preventing venous thromboembolic disease in patients undergoing elective hip and knee arthroplasty guideline.** *J Am Acad Orthop Surg.* 2011;19(Dec):768–776.

40. Bergqvist D, Arcelus JI, Felicissimo P. **Evaluation of the duration of thromboembolic prophylaxis after high-risk orthopaedic surgery: The ETHOS observational study.** *Thromb Haemost.* 2012;107(2):270–279.

41. Gao Y, Long A, Xie Z, et al. **The compliance of thromboprophylaxis affects the risk of venous thromboembolism in patients undergoing hip fracture surgery.** *Springerplus.* 2016;5(1):1362.

42. Wilke T, Müller S. **Nonadherence in outpatient thromboprophylaxis after major orthopedic surgery: a systematic review.** *Expert Rev Pharmacoecon Outcomes Res.* 2010;10(6):691–700.

43. OECD. **OECD Health Data 2009 – comparing health statistics across OECD countries - OECD.** Available at: http://www.oecd.org/health/oecdhealthdata2009comparinghealthstatisticsacrossoecdcountries.htm. Accessed January 15, 2018.

44. Friedman RJ. **Novel oral anticoagulants for VTE prevention in orthopedic surgery: overview of phase 3 trials.** *Orthopedics.* 2011;34(10):795–804.

## SUPPLEMENTAL MATERIAL TO CHAPTER 8

**Appendix 1.** Bivariate logistic regression assessing surgeon and chemical prophylaxis (n=637)

**Appendix 2.** Bivariate logistic regression assessing risk factors for 90-days symptomatic VTE event (n=637)

Supplemental material can be consulted online per the website of the journal and/or publisher.

CHAPTER

09

# COMPLICATIONS AND REOPERATIONS AFTER SURGERY FOR 647 PATIENTS WITH SPINE METASTATIC DISEASE

Nuno R. Paulino Pereira, Paul T. Ogink, Olivier Q. Groot, Marco L. Ferrone, Francis J. Hornicek, Cornelis N. van Dijk, Jos A.M. Bramer, Joseph H. Schwab

*The Spine Journal 2019;19:144-156*

# ABSTRACT

**Background**

Postoperative morbidity may offset the potential benefits of surgical treatment for spine metastatic disease; hence, risk factors for postoperative complications and reoperations should be taken into considerations during surgical decision-making. In addition, it remains unknown whether complications and reoperations shorten these patients' survival.

**Objectives**

Postoperative morbidity may offset the potential benefits of surgical treatment for spine metastatic disease; hence, risk factors for postoperative complications and reoperations should be taken into considerations during surgical decision-making. In addition, it remains unknown whether complications and reoperations shorten these patients' survival.

**Design**

Retrospective cohort study.

**Methods**

A retrospective study identified 647 patients 18 years and older who had surgery for metastatic disease in the spine between January 2002 and January 2014 in one of two affiliated tertiary care centers. The primary outcomes were complications within 30 days after surgery and reoperations until final follow-up or death. We used multivariate logistic regression to identify risk factors for 30-day complications and reoperations. We used the Cox regression analysis to assess the effect of postoperative complications and reoperations on survival.

**Results**

From 647 included patients, 205 (32%) had a complication within 30 days. The following variables were independently associated with 30-day complications: lower albumin levels (OR: 0.69, 95% CI: 0.49 - 0.96, p=0.021), additional comorbidities (OR: 1.42, 95% CI: 1.00 - 2.01, p=0.048), pathologic fracture (OR: 1.41, 95% CI: 0.97 - 2.05, p=0.031), 3 or more spine levels operated upon (OR: 1.64, 95% CI: 1.02 - 2.64, p=0.027), and combined surgical approach (OR: 2.44, 95% CI: 1.06 - 5.60, p=0.036). One hundred and fifteen patients (18%) had at least one reoperation after the initial surgery; prior radiotherapy (OR: 1.56, 95% CI: 1.07 - 2.29, p=0.021) to the spinal tumor was independently associated with reoperation. 30-day complications were associated with worse survival (Hazard Ratio [HR] 1.40, 95% CI: 1.17 - 1.68; p < 0.001), and reoperation was not significantly associated with worse survival (HR 0.80, 95% CI: 0.09 - 1.00; p=0.054). Neurologic status worsened in 42 (6.7%), remained stable in

445 (71%), and improved in 140 (22%) patients after surgery.

**Conclusion**

Three or more spine levels operated upon, and prior radiotherapy should prompt consideration of a pre-operative plastic surgery consultation regarding soft tissue coverage. Furthermore, if time allows, aggressive nutritional supplementation should be considered for patient with low preoperative serum albumin levels. Surgeons should be aware of the increase in complications in patients presenting with pathologic fracture, undergoing a combined approach, and with any additional preoperative comorbidities. Importantly, 30-day complications led were associated with to worsened survival.

# INTRODUCTION

The incidence of bone metastatic disease is increasing as the population ages and patients with cancer survive longer.[1] The spine is the most common site for bony metastases.[2] The goal of operative treatment for spine metastatic disease is to repair neurologic deficits, alleviate pain symptoms, and maintain—or improve—quality of life during the last phase of life.[3,4] Complications or surgical re-interventions may offset the potential benefits of surgical treatment; hence, risk factors for postoperative complications and reoperations should be taken into account during the surgical decision-making.

Numerous studies report on complications after surgery for spine metastatic disease, with rates ranging from 20 to 47%;[5–11] however, many studies are vague in their definition of complications, and study cohort sizes are often relatively small with few outcomes and therefore not able to statistically detect all risk factors.[10–12] In addition, it remains unclear—and difficult—to assess the impact of complications and reoperations on the patients' survival.

With a relatively large cohort, our primary study aim was to describe and identify factors associated with 30-day complications or reoperations. Second, we sought to assess the effect of 30-day complications or reoperations on the patients' postoperative survival. Third, we describe neurologic changes after surgery.

# METHODS

### Study Design and Patients

This retrospective cohort study was approved by our institutional review board. We included patients 18 years and older who underwent operative treatment for metastatic disease in the cervical, thoracic, or lumbar spine between January 2002 and January 2014 in one of two affiliated tertiary care centers. Patients with metastases to the spine from hematologic malignancies–i.e., multiple myeloma and lymphoma—were also included. We excluded patients who presented for revision surgery, or patients who either had stereotactic radiosurgery, vertebroplasty, or kyphoplasty as only procedure. In case a patient underwent multiple surgical procedures for spine metastatic disease, we only included the first procedure to not violate the statistical rule of independence.

We identified 1,330 potentially eligible patients through the ICD-9 code for pathologic fracture (733.13), and an additional 796 patients by a computerized word search in operative reports of our oncology database (containing data for 52,476 patients). We manually screened medical records of the 2,126 potentially eligible patients, and 647 patients met eligibility criteria.

The surgeon decided on the surgical approach by accounting for the patient's estimated survival, neurologic deficits, level of pain, and spinal stability. Patients were followed up at 2 weeks, 6 weeks, and 3 months postoperatively, and subsequently every 3-months until death.

### Outcomes and Explanatory Variables

Our primary outcome measures were complications within 30 days after surgery and reoperations until final follow-up (or death). We graded complications according to the Clavien-Dindo classification:[13] grade I complications were notable postoperative deviations that did not require pharmacological treatment (e.g. conservatively treated pneumothorax); grade II complications were postoperative deviations that required pharmacological treatment –except for commonly used postoperative medications [e.g. analgesics] or blood transfusion; grade III complications were postoperative deviations that required a surgical, endoscopic, or radiologic intervention; grade IV complications were postoperative life-threatening deviations that warranted ICU admission; grade V complications were postoperative deviations resulting in the death of a patient. We dichotomized complications based on the Clavien-Dindo classification into minor complications (grade I or II) and major complications (grade III, IV, or V). We defined reoperations as unplanned surgical reinterventions to the spine directly related to the initial surgery.

Survival was the secondary outcome, defined as death from any cause –as we expected that the majority of deaths were related to metastatic disease in these terminal patients. We screened medical records and the Social Security Death Index (SSDI) to determine survival at the final follow-up moment (i.e. October 4$^{th}$ 2016).[14]

We extracted the following explanatory factors from patients' medical records: age, gender, body mass index (in kg/m$^2$), comorbidity status, Eastern Cooperative Oncology Group (ECOG) performance status, cancer type, number of spine metastatic lesions (excluding sacrum), bone metastases outside the spine, visceral metastases (lung or liver), brain metastases, time between neurologic deficits and surgery (none, less than 14 days, or 14 days or more), location of lesion(s), pathologic fracture, preoperative back pain, prior radiotherapy to the spinal tumor, prior systemic therapy for the cancer diagnosis (all nonsurgical and non-radiotherapeutic adjuvants [chemotherapy, immunotherapy, hormone therapy, and metabolic therapy]), neurologic status before surgery and at discharge using the American Spinal Injury Association (ASIA) impairment scale, hospital, type of surgery and approach, number of spine levels operated upon, type of bone graft(s), type of instrumentation, duration of surgery (in minutes), duration of hospital stay (in days), and estimated blood loss (in milliliters [mL]). Furthermore, we collected the following preoperative laboratory values that were closest to the date of surgery with a maximum of 7 days: hemoglobin levels (g/dL), white blood cell count (10$^3$/ L), creatinine levels (mg/dL), calcium levels (mg/dL), platelet count (10$^3$/mm$^3$), red blood cell count (10$^6$/mm$^3$), albumin levels (g/dL), and lymphocyte count (10$^9$/L).

We used a modified Charlson comorbidity index to grade comorbidity status; this index classifies 12 comorbidities that are associated with 10-year mortality (e.g., diabetes, congestive heart failure).[15] We used a previously reported ICD-9-code-based algorithm to calculate the Charlson comorbidity index in our cohort.[16] We classified comorbidities other than the cancer as presence of additional comorbidities. We categorized cancer type into two groups based on the expected survival, as suggested by Katagiri et al.[17]: those with relatively good prognosis cancers (i.e. lymphoma, multiple myeloma, breast cancer, kidney cancer, prostate cancer, or thyroid cancer), and those with relatively poor prognosis cancers (i.e. lung cancer, colon cancer, rectal cancer, bladder cancer, esophageal cancer, liver cancer, melanoma, gastric cancer, or other cancers). We categorized neurologic status into complete impairment (ASIA grade A), incomplete impairment (ASIA grade B, C, or D), and normal neurologic status (ASIA grade E).[18]

### Statistical Analysis

Categorical variables are described with frequencies and percentages, and continuous variables with medians and interquartile ranges (IQR) as histograms suggested non-normal distributions. We used multivariate logistic regression to identify potential risk factors for 30-day complications –retaining variables with a P-value below 0.10 in bivariate logistic regression. We used multivariate cox regression analysis –retaining variables with a P-value below 0.10 in bivariate cox regression— to identify risk factors for reoperations and used the date of the first reoperation for the time to event analysis. We used Cox regression analysis to assess the effect of postoperative complications and reoperations on survival.

To retain all cases for multivariate analysis we used multiple imputation to estimate missing data. With multiple imputation, the statistical software multiplies the existing dataset multiple times (40 in our case) and substitutes missing values based on all other variables accounting for uncertainty. The statistical software estimated missing values for preoperative albumin levels in 88 patients (14%), preoperative lymphocyte count in 89 patients (14%), and preoperative red blood cell count in 12 patients (2%). We considered P-values less than 0.05 to be significant and used STATA 13 (StataCorp LP, College Station, TX, USA) for statistical analyses.

# RESULTS

Among 647 included patients the median age was 60 years (IQR 52-68), 375 (58%) were men, and the median Charlson comorbidity index score was 6 (IQR 6-8) (Table 1). Two hundred twenty-one patients (34%) had prior radiotherapy to the spine tumor, and 367 (57%) had prior systemic therapy. At presentation, 399 (62%) patients had a pathologic fracture, 215 (33%) had visceral metastases, and 72 (11%) had brain metastases. The thoracic spine was the most common tumor location (379 cases [59%]), and patients most presented with lung cancer (115 cases [18%]) and kidney cancer (81 cases [13%]) (Table 2). Surgical treatments were corpectomy with stabilization (313 cases [48%]), decompression with stabilization (230 cases [36%]), decompression alone (84 cases [13%]), and stabilization alone (20 cases [3.0%]). The median postoperative survival was 305 days (IQR 95 – 957, range 3 - 4823).

### 30-day Complications and Risk Factors

Two hundred and five patients (32%) had a 30-day complication, from which 130 (20%) had a minor complication as the most severe outcome, and 75 (12%) a major complication (Table 3). Systemic infections were encountered in 131 patients (20%), surgical site complications in 83 (13%), thromboembolisms in 49 (6.2%), pulmonary morbidities in 13 (2.0%), cardiac morbidities in 7 (1.0%), gastrointestinal morbidities in 2 (0.5%), and other morbidities in 2 patients (0.5%). Eighteen patients (2.8%) died within 30 days of operative treatment due to the following complications: 6 (0.9%) from pneumonia, 2 (0.3%) from sepsis, 2 (0.3%) from respiratory distress after pulmonary embolism, and 1 (0.2%) from multi-organ failure. Six (0.9%) patients died with a present postoperative complication that did not seem to contribute to the death, and one patient (0.2%) never regained consciousness after surgery without a clear complication focus.

The following factors were associated with 30-day complications in bivariate analysis: lower albumin levels (Odds Ratio [OR]: 0.56, 95% Confidence Interval [CI]: 0.41 – 0.76, p < 0.001), additional comorbidities (OR: 1.49, 95% CI: 1.07 - 2.08. p=0.018), less than 14 days between neurologic deficits

**Table 1.** Baseline characteristics

| Variables | All patients (n=647) |
|---|---|
| | *Median (IQR)* |
| Age (years) | 60 (52–68) |
| Modified Charlson comorbidity index | 6 (6–8) |
| Estimated blood loss (mL)[a] | 715 (400–1500) |
| Duration surgery (minutes)[a] | 385 (286–494) |
| Preoperative laboratory value[a] | |
| Hemoglobin levels (g/dL) | 11 (10–13) |
| White blood cell count ($10^3$/L) | 11 (7.5–14) |
| Creatinine levels (mg/dL) | 0.77 (0.61–0.93) |
| Calcium levels (mg/dL) | 8.7 (8.0–9.3) |
| Platelet count ($10^3$/mm$^3$) | 242 (179–325) |
| Red blood cell count ($10^6$/mm$^3$) | 3.8 (3.4–4.2) |
| Albumin levels (g/dL) | 3.8 (3.4–4.2) |
| Lymphocyte count ($10^9$/L) | 0.92 (0.59–1.5) |
| | *Number (%)* |
| Men | 375 (58) |
| Body mass index (in kg/m$^2$) [a] | |
| < 18.5 | 19 (3.0) |
| 18.5–30 | 402 (72) |
| >30 | 136 (24) |
| Hospital | |
| Hospital 1 | 368 (57) |
| Hospital 2 | 279 (43) |
| ECOG performance status[a] | |
| 0–2 | 347 (79) |
| 3–4 | 93 (21) |
| Preoperative back pain | 548 (85) |
| Additional comorbidities[b] | 300 (46) |
| Time between neurologic deficits and surgery for spine metastatic disease | |
| None | 341 (53) |
| <14 days | 179 (28) |
| ≥14 days | 127 (20) |
| Location of lesion treated for | |
| Cervical | 87 (13) |
| Thoracic | 379 (59) |
| Lumbar | 142 (22) |
| Combined | 39 (6.0) |
| Prognosis of cancer type[c] | |
| Good prognosis | 429 (66) |
| Poor prognosis | 218 (34) |
| Pathologic fracture | 399 (62) |
| Visceral and/or brain metastases | |
| Visceral—lung or liver | 172 (27) |
| Brain | 29 (4.5) |
| Visceral and brain | 43 (6.6) |

*Continued on next page*

| | |
|---|---|
| Number of spine metastatic lesions (excluding sacrum) | |
| 1 | 176 (20) |
| 2 | 98 (15) |
| ≥3 | 373 (58) |
| Bone metastases outside the spine | 341 (53) |
| Prior radiotherapy to the spinal tumor | 221 (34) |
| Prior systemic therapy for the cancer diagnosis | 367 (57) |
| ***Operative variables*** | |
| Type of surgery | |
| Corpectomy with stabilization | 313 (48) |
| Decompression with stabilization | 230 (36) |
| Decompression alone | 84 (13) |
| Stabilization alone | 20 (3.0) |
| Surgical approach | |
| Posterior | 551 (85) |
| Anterior | 68 (11) |
| Combined | 28 (4.0) |
| Number of spine levels operated upon | |
| 1 | 403 (62) |
| 2 | 118 (18) |
| ≥3 | 126 (19) |
| Implants[d] | |
| Allograft | 359 (55) |
| Autograft | 192 (30) |
| Cage | 188 (29) |
| Cement | 103 (16) |
| Plate | 56 (9.0) |

*IQR=interquartile range; mL=milliliter; g/dL=gram per deciliter; L=microliter; mg/dL=milligram per deciliter; $mm^3$=cubic millimeter; kg/$m^2$=kilogram per square meter; L=liter; ECOG=Eastern Cooperative Oncology Group.*

[a] *Estimated blood loss was available in 576 cases (89%), hemoglobin levels in 637 cases (98%), white blood cell count in 636 cases (98%), creatinine levels in 626 cases (96%), calcium levels in 592 cases (91%), platelet count in 635 cases (98%), red blood cell count in 635 cases (98%), albumin levels in 559 (86%), lymphocyte count in 558 cases (86%), body mass index in 557 cases (86%), and ECOG in 440 cases (68%)*

[b] *Based on comorbid conditions in Charlson Comorbidity Index.*

[c] *The good prognosis group includes lymphoma, breast cancer, multiple myeloma, kidney cancer, prostate cancer, and thyroid cancer. The poor prognosis group includes lung cancer, colon cancer, rectal cancer, bladder cancer, esophageal cancer, liver cancer, melanoma, gastric cancer, and other cancers.*

*d Implants used are not mutually exclusive.*

and surgery (OR: 1.63, 95% CI: 1.11 – 2.38, p=0.012), pathologic fracture (OR: 1.47, 95% CI: 1.04 – 2.09, p=0.029), anterior surgical approach (OR: 0.54, 95% CI: 0.29 – 1.00, p=0.048, and 3 or more spine levels operated upon (OR: 1.63, 95% CI: 1.07 – 2.47, p=0.022) (Appendix 1). In multivariate analysis lower albumin levels (OR: 0.69, 95% CI: 0.49 – 0.96, p=0.021), additional comorbidities (OR: 1.42, 95% CI: 1.00 – 2.01, p=0.048), pathologic fracture (OR: 1.41, 95% CI: 0.97 – 2.05, p=0.031), 3 or more spine levels operated upon (OR: 1.64, 95% CI: 1.02 – 2.64, p=0.027), and combined surgical approach (OR: 2.44, 95% CI: 1.06 – 5.60, p=0.036) were independently associated with 30-day complications (Table 4).

**Table 2.** Primary tumor type (n=647)

| *Primary tumor type* | *Number (%)* |
|---|---|
| Lung | 115 (18) |
| Kidney | 81 (13) |
| Breast | 77 (12) |
| Multiple myeloma | 71 (11) |
| Prostate | 57 (8.8) |
| Melanoma | 27 (4.2) |
| Colorectal | 25 (3.9) |
| Neuroendocrine | 21 (3.2) |
| Sarcomatous | 21 (3.2) |
| Unknown primary | 20 (3.1) |
| Lymphoma | 19 (2.9) |
| Head and neck | 17 (2.6) |
| Thyroid | 16 (2.5) |
| Hepatocellular | 13 (2.0) |
| Esophageal | 12 (1.9) |
| Endometrial | 11 (1.7) |
| Salivary carcinoma | 6 (0.9) |
| Other* | 38 (5.9) |

**Other cancer types were bladder five cases (0.8%), adenocarcinoma in five cases (0.8%), germ cell in five cases (0.8%), pancreatic in five (0.8%), ovarian in three cases (0.5%), testicular in three cases (0.5%), penile in three cases (0.5%), cholangiocarcinoma in three cases (0.5%), gastric in two cases (0.3%), adrenal in two cases (0.3%), blue round cell in one case (0.2%), skin in one case (0.2%), and leukemia in one case (0.2%).*

### Reoperations and Risk Factors

One hundred and fifteen patients (18%) had at least one reoperation after the initial surgery (Figure 1). The most common reasons for reoperation were recurrent tumor in 45 (7.0%), wound infection in 42 (6.5%), and wound dehiscence in 28 patients (4.3%) (Table 5). Prior radiotherapy to the spinal tumor was the only factor that was associated with reoperation in bivariate (OR: 1.68, 95% CI: 1.16 – 2.43, p=0.006) (Appendix 2) and multivariate analysis (OR: 1.56, 95% CI: 1.07 – 2.29, p=0.021) (Table 6).

### Effect of Complications and Reoperations on Patients' Survival

A bivariate analysis showed that 30-day complications were associated with worse survival (Hazard Ratio [HR] 1.40, 95% CI: 1.17 – 1.68; p < 0.001) (Figure 2). Compared to patients who had no complications, both patients with minor complications (HR 1.29, 95% CI: 1.04 – 1.60; p=0.019), and patients with major complications had worse survival (HR 1.63, 95% CI: 1.25 – 2.13; p < 0.001) (Figure 3). Reoperation was not significantly associated with survival (HR 0.80, 95% CI: 0.64 – 1.00; p=0.054) (Figure 4).

Table 3. 30-Days complications (n=647)

| | *Number (%)* |
|---|---|
| At least one complication | 205 (32) |
| Major | 130 (20) |
| Minor | 75 (12) |
| Number of complications | |
| 1 | 122 (19) |
| 2 | 63 (9.6) |
| 3 or more | 21 (3.2) |
| Clavien-Dindo classification | |
| 1. No need for further intervention | 8 (3.9) |
| 2. Requiring pharmacologic treatment | 122 (59) |
| 3. Requiring surgery/endoscopy | 50 (24) |
| 4. Life-threatening complication | 8 (3.8) |
| 5. Death due to complication | 17 (8.3) |
| Complication types | |
| Surgical site complication | 83 (13) |
| Wound infection | 48 (7.4) |
| Wound dehiscence | 25 (3.9) |
| Hematoma at surgical site | 2 (0.3) |
| Epidural hematoma | 2 (0.3) |
| Fractured vertebra | 2 (0.3) |
| Hardware displacement | 1 (0.2) |
| Extravasation of cement | 1 (0.2) |
| Active wound bleeding | 1 (0.2) |
| Spinal fluid leak | 1 (0.2) |
| Systemic infection | 131 (20) |
| Urinary tract infection | 58 (9.0) |
| Pneumonia | 56 (8.7) |
| Sepsis | 16 (2.5) |
| Viral gastrointestinal infection | 1 (0.2) |
| Thromboembolism | 49 (6.2) |
| DVT | 29 (4.5) |
| PE | 23 (3.6) |
| Pulmonary morbidity | 13 (2.0) |
| Pneumothorax | 9 (1.4) |
| Respiratory failure | 3 (0.5) |
| ARDS | 1 (0.2) |
| Cardial morbidity | 7 (1.0) |
| Myocardial infarct | 5 (0.8) |
| Atrial fibrillation | 2 (0.3) |
| Gastrointestinal morbidity | 2 (0.5) |
| Ileus | 1 (0.2) |
| Gastric ulcer | 1 (0.2) |
| Other | 2 (0.5) |
| Cerebellar infarct | 1 (0.2) |
| Never regained consciousness after surgery | 1 (0.2) |

*PE=pulmonary embolism; DVT=deep vein thrombosis; ARDS=acute respiratory distress syndrome.*

**Table 4.** Multivariate logistic regression assessing risk factors for 30-day complication after multiple imputation (40 imputations; n=647)

| *Variables* | *Odds ratio (95% CI)* | *Standard error* | *P-value* |
|---|---|---|---|
| Albumin levels (g/dL)* | 0.69 (0.49–0.96) | 0.12 | **0.021** |
| Lymphocyte count ($10^9$/L)* | 0.91 (0.76–1.10) | 0.09 | 0.326 |
| Additional comorbidities† | 1.42 (1.00–2.01) | 0.25 | **0.048** |
| Time between neurologic deficits and surgery for spine metastatic disease | | | |
| None | *Reference value* | | |
| <14 days | 1.28 (0.85–1.92) | 0.27 | 0.117 |
| ≥14 days | 0.82 (0.51–1.33) | 0.20 | 0.540 |
| Location of lesion treated for | | | |
| Cervical | 0.69 (0.39–1.21) | 0.20 | 0.149 |
| Thoracic | *Reference value* | | |
| Lumbar | 1.24 (0.80–1.93) | 0.28 | 0.441 |
| Combined | 0.64 (0.29–1.44) | 0.26 | 0.321 |
| Pathologic fracture | 1.41 (0.97–2.05) | 0.27 | **0.031** |
| Number of spine levels operated upon | | | |
| 1 | *Reference value* | | |
| 2 | 1.25 (0.78–2.01) | 0.30 | 0.244 |
| ≥3 | 1.64 (1.02–2.64) | 0.40 | **0.027** |
| Surgical approach | | | |
| Posterior | *Reference value* | | |
| Anterior | 0.63 (0.32–1.21) | 0.21 | 0.166 |
| Combined | 2.44 (1.06–5.60) | 1.03 | **0.036** |

*g/dL=gram per deciliter; L=liter.; CI=confidence intervals.* **Bold** *indicates significance (P<0.05).*
** Albumin levels were available in 559 cases (86%) and lymphocyte count in 558 cases (86%).*
*† Based on comorbid conditions in Charlson comorbidity index.*

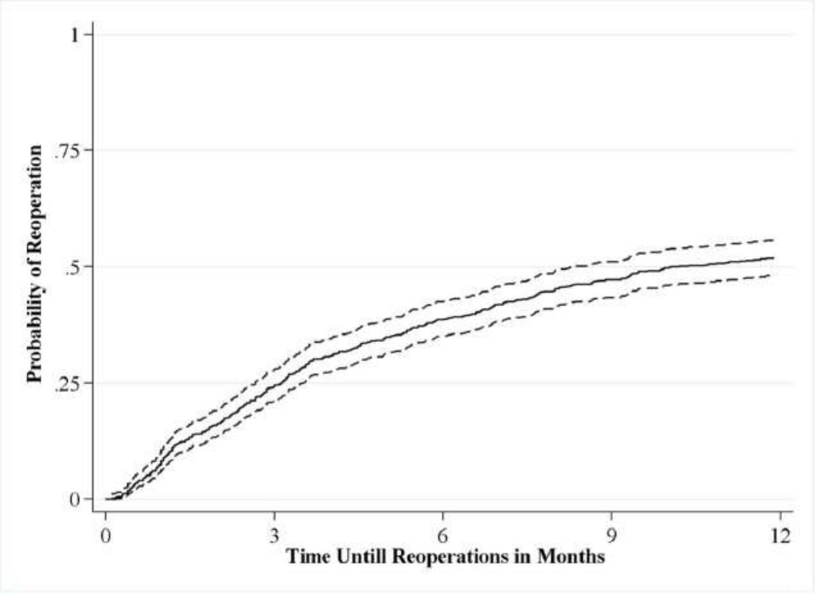


**Figure 1.** Reverse Kaplan-Meier plot showing the risk of reoperation over time (this figure only accounts for first reoperations).

**Table 5.** Overview of the reoperations (n=647)

| | *Number (%)* |
|---|---|
| At least one reoperation | 115 (18) |
| Number of reoperations | |
| 1 | 81 (13) |
| 2 | 22 (3.4) |
| 3 or more | 12 (1.9) |
| Reason for reoperations* | |
| Surgical site | |
| Wound infection | 42 (6.5) |
| Wound dehiscence | 28 (4.3) |
| Hematoma | 5 (0.8) |
| Seroma | 4 (0.6) |
| Epidural hematoma | 4 (0.6) |
| Necrotic tissue | 1 (0.2) |
| Epidural abscess | 1 (0.2) |
| Cerebrospinal fluid leak | 1 (0.2) |
| Paraspinal abscess | 1 (0.2) |
| Hardware | |
| Hardware failure | 10 (1.5) |
| Exposed hardware | 4 (0.6) |
| Hardware displacement | 3 (0.5) |
| Painful hardware | 3 (0.5) |
| Cement extravasation | 1 (0.2) |
| Tumor | |
| Recurrent tumor | 45 (7.0) |
| Remaining tumor after initial surgery | 6 (0.9) |
| Other | |
| Failed anterior fusion at initial surgery | 5 (0.8) |
| Non-union | 2 (0.3) |
| Pseudo meningocele | 2 (0.3) |
| Muscle flap transposition | 1 (0.2) |
| Placement tracheostomy | 1 (0.2) |
| Oropharyngeal fistula | 1 (0.2) |
| Correction of spinal alignment | 1 (0.2) |
| Further posterior stabilization | 1 (0.2) |
| Removal of VAC dressing | 1 (0.2) |

*VAC=Vacuum-assisted closure.*
**Not mutually exclusive.*

**Table 6.** Multivariate cox regression analysis assessing risk factors for reoperation after multiple imputation (40 imputations; n=647)

| *Variables* | *Hazard ratio (95% CI)* | *Standard error* | *P-value* |
|---|---|---|---|
| Red blood cell count ($10^6$/mm$^3$)* | 0.79 (0.57–1.10) | 0.13 | 0.158 |
| Prior radiotherapy to the spinal tumor | 1.56 (1.07–2.29) | 0.3 | **0.021** |
| Prior systemic therapy for the cancer diagnosis | 1.20 (0.81–1.77) | 0.24 | 0.373 |

*CI=confidence interval; mm$^3$=cubic millimeter.* **Bold** *indicates significance ($P<0.05$).*
**Preoperative red blood cell count was available in 635 cases (98%).*

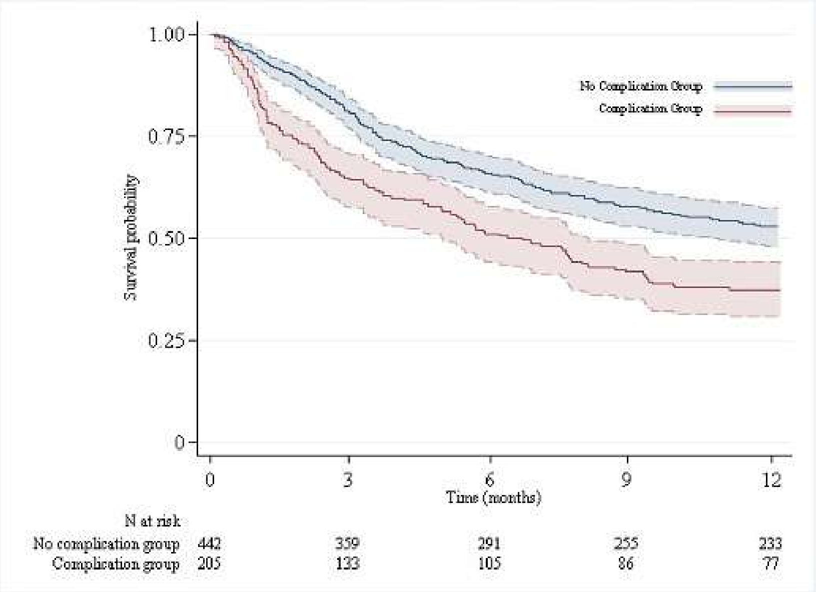


**Figure 2.** Kaplan-Meier plot showing the probability of survival with 95% confidence interval for patients with (red line) and without 30-day complications (blue line). Bivariate Cox regression analysis showed a significant difference in survival (HR 1.40, 95%CI (1.17–1.68), p<0.001).

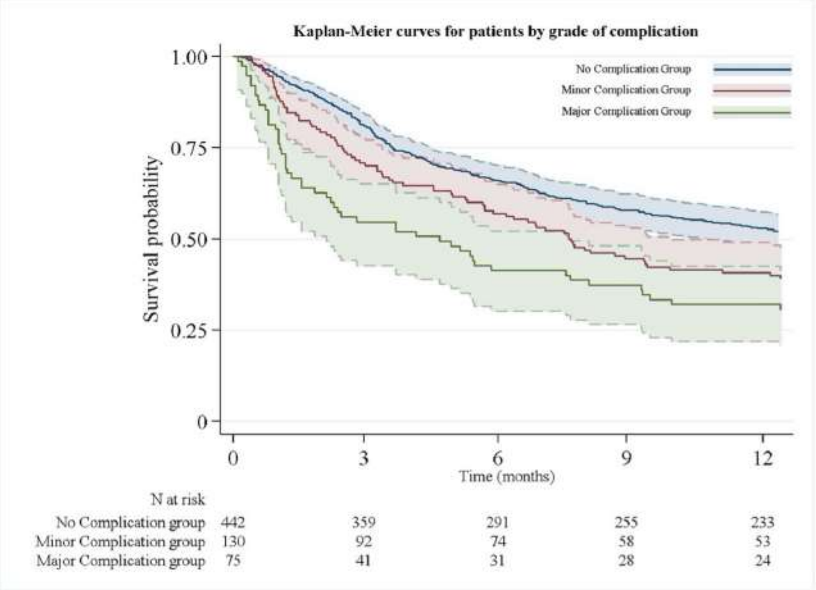


**Figure 3.** Kaplan-Meier plot showing the probability of survival with 95% confidence interval for patients without 30-day complications (blue line), patients with minor 30-day complications (red line), and patients with major 30-day complication (green line). Bivariate Cox regression analysis showed a significant difference in survival between no complications and minor complication (HR 1.29, 95% CI (1.04–1.60), p=.019), and between no complications and major complications (HR 1.63, 95% CI (1.25–2.13), p<.001).

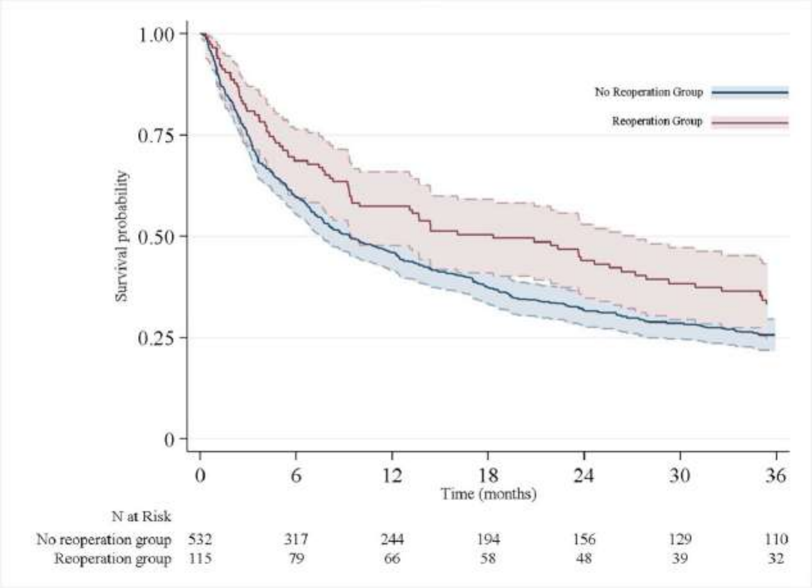


**Figure 4.** Kaplan-Meier plot showing the probability of survival with 95% confidence interval for patients without a reoperation (blue line) and patients with reoperation (red line). Bivariate Cox regression analysis showed no significant difference in survival (HR 0.80, 95% CI (0.64–1.00), p=.054).

**Table 7.** Neurologic status before surgery and at discharge (n=647)

| | | Discharge status† | | |
|---|---|---|---|---|
| | | Complete | Incomplete | Normal |
| Preoperative status* | Complete | 2 | 4 | 1 |
| | Incomplete | 4 | 151 | 135 |
| | Normal | 0 | 38 | 292 |

*Based on ASIA impairment scale: Complete: Grade A, Incomplete: Grade B/C/D, Normal: Grade E.*
**Preoperative neurologic status available in 637 cases (98%) and discharge neurologic status in 636 cases (98%).*

### Neurologic Changes

Before surgery, 7 patients (1.1%) presented with complete neurologic impairment (ASIA grade A), 299 (47%) with incomplete neurologic impairment (ASIA grade B, C, or D), and 331 (52%) with normal neurologic status (ASIA grade E). At discharge, 6 patients (0.9%) had complete impairment (ASIA grade A), 195 (31%) had incomplete impairment (ASIA grade B, C, or D), and 435 (68%) had a normal neurologic status (ASIA grade E). Neurologic status worsened in 42 (6.7%), remained equal in 445 (71%), and improved in 140 (22%) patients after surgery (Table 7).

# DISCUSSION

Despite numerous studies on complications and reoperation after surgery for spine metastatic disease, it remains unclear what risk factors attribute to these outcomes.[5–11] In the present study we (1) sought to identify patients more at risk for postoperative complications and reoperations, (2) assess the effect of complications and reoperations on survival, and (3) describe neurologic changes after surgery. Lower preoperative albumin levels, additional comorbidities, pathologic fracture, more than 3 spine levels operated on, and combined surgical approach were independent risk factors for 30-day complications; prior radiotherapy was the only risk factor for having a reoperation after the initial surgery. Having a 30-day complication resulted in decreased survival–both for patients with minor and major complications–, and reoperations did not significantly affect postoperative survival rates. The patients' neurologic status remained equal in most patients, and surgery was able to improve the neurologic status in about 1/5 of all patients.

This study has several limitations. First, we may have missed complications and reoperations that occurred outside our facilities, which may have led to an underestimation of true complication and reoperation rates. However, only 1.7% of the patients were lost-to follow-up within 30-day, and that is why we believe the 30-day time frame to be safe. Second, the retrospective nature of this study may have led to selection bias; choice of surgical treatment was dependent on individual providers and may lack uniformity. Third, there is a large variation in follow-up and survival of patients, inherent to a group of patients with advanced cancer. This may lead to difficulties applying the conclusions of our study to individual patients. We tried to offset this for reoperations by using a Cox regression analysis which accounts for this variation. Fourth, certain variables in our study suffer from interobserver variability (e.g., neurologic status, performance status). We tried to account for this by categorizing these variables using cut off point representing meaningful qualitative differences. Fifth, the use of diagnosis codes to identify patients and for constructing the modified Charlson Comorbidity Index may have led to inaccuracies due to miscoding in clinical practice. Finally, our study population consists of many different cancer types resulting in heterogeneity. However, in bivariate analysis none of the 5 most prevalent cancer types were associated with 30-day complications or reoperations (Appendix 3).

Despite these limitations, we believe our study represents a large cohort of patients with spine metastatic disease adequately powered to identify risk factors for complications and reoperations. A study by Arrigo et al.[12] (n=200) found similar complication rates in patients who underwent surgery for spine metastatic disease (34% vs. 32% in our study). The Charlson Comorbidity Index was the only independent risk factor for overall complications, although they may have been underpowered to detect other risk factors. Importantly, and contrary to our study, they found a high variation in complication rates per cancer type. A recent study by Schoenfeld et al.[19] assessed the predictive accuracy of the New England Spinal Metastasis Score (NESMS) for 30-day morbidity. The NESMS is a prognostic tool that aims to prognosticate 1-year survival after surgery for spine metastatic disease and is based on ambulatory status, preoperative albumin levels, and the modified Bauer score. They concluded that the NESMS was a significant predictor for 30-day mortality, major systemic complications, and death after major complications. A novel and important finding of their study was that factors associated with worse survival were also predictive for more complications –which is in line with our study. Although our study focused on all 30-day complications we also found preoperative albumin levels to be associated with complications.

The overall 30-day complication rate (32%) and major complication rate (12%) encountered in our hospitals, are comparable with overall (ranging from 20 to 47%)[5–11] and major complication rates in the literature (13.8 to 27%).[6,8,10,11] An explanation for our relatively low major complication rate may be due to different definitions of major complication. Unfortunately, most articles have unclear definitions for major complications, making it difficult to compare. Another explanation may be that the two high-volume tertiary centers have lower major complication rates due to more experience with this type of complicated surgery and care.

Lower albumin levels have long been associated with increased complication rates after oncologic surgery as well as after other types of orthopedic surgery.[20–27] Although serum albumin is usually seen as a marker for nutritional status, some advocate that hypoalbuminemia is caused by systemic inflammation –and not solely by a lowered nutritional status; the increased demand of protein for acute-phase proteins in systemic inflammation results in a fall in serum albumin levels.[28–30] Similarly, the association of additional comorbidities with 30-day complications is likely a reflection of an overall diminished health status. Other studies have found an association between comorbidities and readmission after surgery for spinal tumors.[31,32] Although surgeons will already be wary of patients' comorbidities, it is useful to reaffirm their important role in surgical decision-making.

Pathologic fracture was identified as an independent risk factor for 30-day complications. Although similar studies did not find pathologic fracture to be associated with complications there is recognition that it negatively impacts survival. Bauer et al.[33] found pathologic fracture to be an independent marker for worse 1-year survival. Behnk et al.[34] reported that surgical site infection, acute myocard infarction, pneumonia, and pulmonary embolism were independently associated with increased same-admission mortality in patients with pathological fractures. Combined with our findings this suggests the gravity of presenting with pathological fracture by being both associated with complications and with subsequent worsened survival.

Three or more spine levels operated upon was another independent risk factor for 30-day complications. Although similar oncologic studies did not specifically look for or find this association, multilevel surgery has been associated with increased complications after degenerative and traumatic spinal surgery.[35–37] Among the 126 patients with 3 or more levels operated upon, most common postoperative complications were wound infection (10/126, 7.9%) and wound dehiscence (5/126, 4.0%); we therefore believe that larger incisions (required for multilevel surgery) are most responsible for increased complication rates. Extra attention to wound closure and postoperative wound care is warranted in these patients to prevent (wound) complications.[38]

Combined approach was not used frequently in our study (4%) but was independently associated with 30-days complications. Similarly, in an early study on combined approach Sundaresan et al.[39] reported a high incidence of surgical complications (48%). Furthermore, both Street et al.[8] and Shehadi et al.[11] found higher estimated blood losses compared to other approaches; an association we did not find in our study (p=0.095). Because the combined approach is utilized when a singular approach does not suffice, surgeons need to be aware of the increased risk of complications.

The reoperation rate in our cohort (18%) is comparable with rates in the literature (10.3 to 47.5%).[5,7–9] Although most (72/115, 63%) initial reoperations occurred within 2 months, some reoperations occurred after 2 years (12/115, 10%); this emphasizes the importance of prognosticating life-expectancy in surgical decision-making, since the choice for a surgical treatment is largely based on a patient's life expectancy. Survival algorithms could aid surgeons in this difficult–yet important—task.[40]

Prior radiotherapy was the only variable associated with more reoperations. One of the side-effects of radiotherapy is delayed wound healing.[41] Fifty-five percent of the reasons for reoperation in the irradiated group were due to wound infections, wound dehiscence, epidural hematoma, or a combination of those compared to 38% in the group without radiotherapy (p=0.089; Fisher's exact test). Ghogawala et al.[42] found a threefold major wound complication rate in patients who had prior radiotherapy compared to patients without prior radiotherapy. Similarly, Sundaresan et al.[9] found that all wound complications requiring reoperation were in the prior irradiated group. Contradictory, Street et al.[8] did not find an association between prior radiotherapy and wound failure, theorizing this is because none of their patients were operated on within 7 days of radiotherapy. A systemic review by Itshayek et al.[43] aimed to find an optimal interval between radiotherapy and surgery to avoid wound complications, and recommended to wait 1 week after radiotherapy; however, the ideal time will likely vary based on patient characteristics.

Patients with a 30-day complication had significantly worse survival than patients without a complication: 28% of the patients with a 30-day complication die within one month, 6.2% in the group without complications (Figure 1). Similarly, Jansson et al.[7] found that systemic complications after surgery for spine metastatic disease often led to premature postoperative death. Other articles studied independent predictors of survival, but did not include postoperative complications as possible predictor.[6,44] Several non-orthopedic oncologic surgery studies did find an association between postoperative complications and worse survival.[45,46] Our results indicate that postoperative complications shorten survival; therefore, it is important to carefully select patients for surgery that are less prone to develop postoperative complications.

Our number of patients with a decline in neurologic status after surgery (6.7%) is within the range of previous studies (0% to 6.9%)[7–11], however the literature on the influence of neurologic status on survival is conflicting. Finkelstein et al.[6] found preoperative neurologic deficit to be an independent risk factor for decreased survival, whereas Jansson et al.[7] did not. Although we did not find a statistically significant effect for neurologic improvement (p=0.146) or neurologic worsening (p=0.380) on survival, the number of patients in our studies who either retained (71%) or improved neurologic status (22%) proves that decompressive surgery for spine metastatic disease reaches its goals in most patients.

# CONCLUSION

Surgery for spine metastatic disease on 3 or more spine levels and prior radiotherapy should prompt consideration of a pre-operative consultation with plastic surgery about soft tissue coverage. Postponing surgery for 1 week after radiotherapy could decrease the risk for wound complications. Furthermore, if time allows, aggressive nutritional supplementation should be considered for patient with low preoperative albumin levels. Although a combined approach may be necessary, surgeons need to acknowledge the possible risk associated with this approach and contemplate its necessity. Additionally, surgeons should be aware of the increase in complications in patients presenting with pathologic fracture and with any additional preoperative comorbidities. Importantly, our study shows that in patients with spine metastatic disease, 30-day complications were associated with ~~to~~ worsened survival.

# REFERENCES

1. Quinn RH, Randall RL, Benevenia J, et al. **Contemporary management of metastatic bone disease: tips and tools of the trade for general practitioners.** *Instr Course Lect.* 2014;63:431–41.

2. Mesfin A, Buchowski JM, Gokaslan ZL, et al. **Management of metastatic cervical spine tumors.** *J Am Acad Orthop Surg.* 2015;23(1):38–46.

3. Fehlings MG, Nater A, Tetreault L, et al. **Survival and clinical outcomes in surgically treated patients with metastatic epidural spinal cord compression: results of the prospective multicenter AOSpine study.** *J Clin Oncol.* 2016;34(3):268–276.

4. Quan GMY, Vital J-M, Aurouer N, et al. **Surgery improves pain, function and quality of life in patients with spinal metastases: a prospective study on 118 patients.** *Eur Spine J.* 2011;20(11):1970–8.

5. Falicov A, Fisher CG, Sparkes J, et al. **Impact of surgical intervention on quality of life in patients with spinal metastases.** *Spine (Phila. Pa. 1976).* 2006;31(24):2849–2856.

6. Finkelstein JA, Zaveri G, Wai E, et al. **A population-based study of surgery for spinal metastases. Survival rates and complications.** *J Bone Joint Surg Br.* 2003;85(7):1045–1050.

7. Jansson KÅ, Bauer HCF. **Survival, complications and outcome in 282 patients operated for neurological deficit due to thoracic or lumbar spinal metastases.** *Eur Spine J.* 2006;15(2):196–202.

8. Street J, Fisher C, Sparkes J, et al. **Single-stage posterolateral vertebrectomy for the management of metastatic disease of the thoracic and lumbar spine: a prospective study of an evolving surgical technique.** *J Spinal Disord Tech.* 2007;20(7):509–520.

9. Sundaresan N, Rothman A, Manhart K, et al. **Surgery for solitary metastases of the spine: rationale and results of treatment.** *Spine (Phila. Pa. 1976).* 2002;27(16):1802–1806.

10. Wise JJ, Fischgrund JS, Herkowitz HN, et al. **Complication, survival rates, and risk factors of surgery for metastatic disease of the spine.** *Spine (Phila. Pa. 1976).* 1999;24(18):1943–1951.

11. Shehadi JA, Sciubba DM, Suk I, et al. **Surgical treatment strategies and outcome in patients with breast cancer metastatic to the spine: a review of 87 patients.** *Eur Spine J.* 2007;16(8):1179–92.

12. Arrigo RT, Kalanithi P, Cheng I, et al. **Charlson score is a robust predictor of 30-day complications following spinal metastasis surgery.** *Spine (Phila. Pa. 1976).* 2011;36(19):E1274-80.

13. Clavien PA, Barkun J, de Oliveira ML, et al. **The Clavien-Dindo classification of surgical complications: five-year experience.** *Ann Surg.* 2009;250(2):187–196.

14. Huntington JT, Butterfield M, Fisher J, et al. **The Social Security Death Index (SSDI) most accurately reflects true survival for older oncology patients.** *Am J Cancer Res.* 2013;3(5):518–522.

15. Charlson ME, Pompei P, Ales KL, et al. **A new method of classifying prognostic comorbidity in longitudinal studies: development and validation.** *J Chronic Dis.* 1987;40(5):373–383.

16. Quan H, Li B, Couris CM, et al. **Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries.** *Am J Epidemiol.* 2011;173(6):676–682.

17. Katagiri H, Okada R, Takagi T, et al. **New prognostic factors and scoring system for patients with skeletal metastasis.** *Cancer Med.* 2014;3(5):1359–1367.

18. Ditunno JFJ, Young W, Donovan WH, et al. **The international standards booklet for neurological and**

functional classification of spinal cord injury. American Spinal Injury Association. *Paraplegia*. 1994;32(2):70–80.

19. Schoenfeld AJ, Le H V, Marjoua Y, et al. Assessing the utility of a clinical prediction score regarding 30-day morbidity and mortality following metastatic spinal surgery: the New England Spinal Metastasis Score (NESMS). *Spine J.* 2016;16(4):482–490.

20. Bohl DD, Shen MR, Kayupov E, et al. Hypoalbuminemia independently predicts surgical site infection, pneumonia, length of stay, and readmission after total joint arthroplasty. *J Arthroplasty*. 2016;31(1):15–21.

21. Dietch ZC, Guidry CA, Davies SW, et al. Hypoalbuminemia is disproportionately associated with adverse outcomes in obese elective surgical patients. *Surg Obes Relat Dis.* 2015;11(4):912–918.

22. Lohsiriwat V, Lohsiriwat D, Boonnuch W, et al. Pre-operative hypoalbuminemia is a major risk factor for postoperative complications following rectal cancer surgery. *World J Gastroenterol.* 2008;14(8):1248–1251.

23. Matsuoka K, Misaki N, Sumitomo S. Preoperative hypoalbuminemia is a risk factor for late bronchopleural fistula after pneumonectomy. *Ann Thorac Cardiovasc Surg.* 2010;16(6):401–405.

24. Uppal S, Al-Niaimi A, Rice LW, et al. Preoperative hypoalbuminemia is an independent predictor of poor perioperative outcomes in women undergoing open surgery for gynecologic malignancies. *Gynecol Oncol.* 2013;131(2):416–422.

25. Asher V, Lee J, Bali A. Preoperative serum albumin is an independent prognostic predictor of survival in ovarian cancer. *Med Oncol.* 2012;29(3):2005–2009.

26. Ataseven B, du Bois A, Reinthaller A, et al. Pre-operative serum albumin is associated with post-operative complication rate and overall survival in patients with epithelial ovarian cancer undergoing cytoreductive surgery. *Gynecol Oncol.* 2015;138(3):560–565.

27. Gibbs J, Cull W, Henderson W, et al. Preoperative serum albumin level as a predictor of operative mortality and morbidity: results from the National VA Surgical Risk Study. *Arch Surg.* 1999;134(1):36–42.

28. Crumley ABC, Stuart RC, McKernan M, et al. Is hypoalbuminemia an independent prognostic factor in patients with gastric cancer? *World J Surg.* 2010;34(10):2393–2398.

29. McMillan DC, Crozier JEM, Canna K, et al. Evaluation of an inflammation-based prognostic score (GPS) in patients undergoing resection for colon and rectal cancer. *Int J Colorectal Dis.* 2007;22(8):881–886.

30. McMillan DC, Watson WS, O'Gorman P, et al. Albumin concentrations are primarily determined by the body cell mass and the systemic inflammatory response in cancer patients with weight loss. *Nutr Cancer*. 2001;39(2):210–213.

31. Karhade AV, Vasudeva VS, Dasenbrock HH, et al. Thirty-day readmission and reoperation after surgery for spinal tumors: a National Surgical Quality Improvement Program analysis. *Neurosurg Focus*. 2016;41(2):E5.

32. Schairer WW, Carrer A, Sing DC, et al. Hospital readmission rates after surgical treatment of primary and metastatic tumors of the spine. *Spine (Phila. Pa. 1976)*. 2014;39(21):1801–1808.

33. Bauer HC, Wedin R. Survival after surgery for spinal and extremity metastases. Prognostication in 241 patients. *Acta Orthop Scand.* 1995;66(2):143–146.

34. Behnke NK, Baker DK, Xu S, et al. Risk factors for same-admission mortality after pathologic fracture secondary to metastatic cancer. *Support care cancer.* 2017;25(2):513–521.

35. Hartig D, Batke J, Dea N, et al. Adverse events in surgically treated cervical spondylopathic myelopathy: a prospective validated observational study. *Spine (Phila. Pa. 1976)*. 2015;40(5):292–298.

36. Leckie S, Yoon ST, Isaacs R, et al. **Perioperative complications of cervical spine surgery: analysis of a prospectively gathered database through the Association for Collaborative Spinal Research.** *Glob spine J.* 2016;6(7):640–649.

37. Lonjon G, Dauzac C, Fourniols E, et al. **Early surgical site infections in adult spinal trauma: a prospective, multicentre study of infection rates and risk factors.** *Orthop Traumatol Surg Res.* 2012;98(7):788–794.

38. Cassir N, De La Rosa S, Melot A, et al. **Risk factors for surgical site infections after neurosurgery: A focus on the postoperative period.** *Am J Infect Control.* 2015;43(12):1288–1291.

39. Sundaresan N, Steinberger AA, Moore F, et al. **Indications and results of combined anterior-posterior approaches for spine tumor surgery.** *J Neurosurg.* 1996;85(3):438–446.

40. Paulino Pereira NR, Janssen SJ, van Dijk E, et al. **Development of a prognostic survival algorithm for patients with metastatic spine disease.** *J Bone Jt Surg Am.* 2016;98(21):1767–1776.

41. Haubner F, Ohmann E, Pohl F, et al. **Wound healing after radiation therapy: review of the literature.** *Radiat Oncol.* 2012;7:162.

42. Ghogawala Z, Mansfield FL, Borges LF. **Spinal radiation before surgical decompression adversely affects outcomes of surgery for symptomatic metastatic spinal cord compression.** *Spine (Phila. Pa. 1976).* 2001;26(7):818–824.

43. Itshayek E, Yamada J, Bilsky M, et al. **Timing of surgery and radiotherapy in the management of metastatic spine disease: a systematic review.** *Int J Oncol.* 2010;36(3):533–544.

44. Lau D, Leach MR, La Marca F, et al. **Independent predictors of survival and the impact of repeat surgery in patients undergoing surgical treatment of spinal metastasis.** *J Neurosurg Spine.* 2012;17(6):565–576.

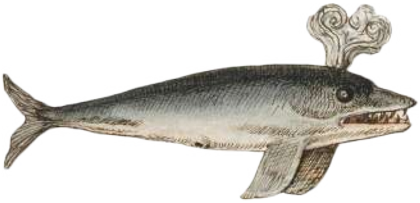
45. Odermatt M, Miskovic D, Flashman K, et al. **Major postoperative complications following elective resection for colorectal cancer decrease long-term survival but not the time to recurrence.** *Color Dis.* 2015;17(2):141–149.

46. Aahlin EK, Olsen F, Uleberg B, et al. **Major postoperative complications are associated with impaired long-term survival after gastro-esophageal and pancreatic cancer surgery: a complete national cohort study.** *BMC Surg.* 2016;16(1):32.

# SUPPLEMENTAL MATERIAL TO CHAPTER 9

**Appendix 1.** Bivariate logistic regression assessing risk factors for 30-day complications
**Appendix 2.** Bivariate cox regression analysis assessing risk factors for reoperation (n=647)
**Appendix 3.** Complications and reoperations for the five most prevalent cancer types

Supplemental material can be consulted online per the website of the journal and/or publisher.



PART IV

# Supplementing Artificial Intelligence Tools

CHAPTER

10

# BODY COMPOSITION PREDICTORS OF MORTALITY IN PATIENTS UNDERGOING SURGERY FOR LONG BONE METASTASES

Olivier Q. Groot*, Michiel E.R. Bongers*, Colleen G. Buckless*, Peter K. Twining, Neal D. Kapoor, Stein J. Janssen, Joseph H. Schwab, Martin Torriani, Miriam A. Bredella

**Joint first authorship*

*Journal of Surgical Oncology (Under review)*

# ABSTRACT

### Background

Although survival of patients with spinal metastases has improved over the last decades due to advances in multi-modal therapy, there are currently no reliable predictors of mortality. Computed tomography (CT) body composition measurements have been recently proposed as biomarkers for survival in patients with and without cancer. Patients with cancer routinely undergo CT for staging or surveillance of therapy and body composition assessed using opportunistic CTs might be used to determine survival in patients with spinal metastases.

### Objectives

To determine the value of body composition measures obtained on opportunistic abdomen CTs to predict 90-day and 1-year mortality in patients with spinal metastases undergoing surgery.

### Design

Retrospective imaging study.

### Methods

Between 2001 and 2016, 196 patients who underwent surgery for spinal metastases at a single tertiary center underwent CT of the abdomen within three months prior to surgery. Quantification of cross-sectional areas (CSA) and CT attenuation in Hounsfield Units (HU) of abdominal subcutaneous adipose tissue (SAT), visceral adipose tissue (VAT), and paraspinous skeletal muscle were performed on CT images at the level of L4 using an in-house automated algorithm under the supervision of a fellowship-trained musculoskeletal radiologist. Sarcopenia was determined by total muscle area (cm2) divided by height squared (m2) with cutoff values of <52.4 cm2/m2 for men and <38.5 cm2/m2 for women. Bivariate and multivariate Cox proportional-hazard analyses were used to determine the associations between body compositions and 90-day and 1-year mortality.

### Results

The median age was 62 years (interquartile range=53-70). Mortality rates for 90-day and 1-year were 24% and 54%, respectively. Sarcopenia and decreased muscle attenuation were associated with increased mortality for both timepoints ($p=0.04$ and $p=0.04$, respectively) after controlling for sex, age, body mass index, Charlson Comorbidity score, primary tumor type, visceral metastases, and duration between diagnosis of spinal metastases and surgery. Decreased SAT area was associated with increased 90-day mortality after controlling for the same covariates ($p<0.01$).

**Conclusion**

Decreased muscle attenuation and sarcopenia are independently associated with an increased risk of 90-day and 1-year mortality for patients surgically treated for spinal metastases, and low SAT CSA is independently associated with increased risk of 90-day mortality. Therefore, body composition measurements could serve as novel biomarkers for prediction of mortality in patients with spinal metastases.

# INTRODUCTION

Long bones are a common site for metastatic disease, and long bone metastases are found in up to 70% of patients with advanced neoplastic disease.[1] Metastases to long bone compromise the structural integrity of the bone and its ability for load-bearing, which can initially lead to painful microfractures followed by pathological fractures, which are associated with a decline in quality of life.[2] Surgical stabilization is often performed for patients with pathological fractures of long bones, but prophylactic stabilization is also regularly considered for patients with known metastatic disease at high risk for a fracture. Due to the incurable nature of metastatic disease, treatment for these patients is primarily performed for palliative measures to maintain or optimize quality of life. For some patients, the benefits of surgery may not outweigh the disadvantages that come with it such as perioperative mortality, postoperative complications, hospitalization, and reoperations.[3,4] Less intensive treatment, such as radiation therapy or minimally invasive stabilization, might be more appropriate for patients with an estimated short survival. Expected survival is thus an important factor in decision making of the most-appropriate therapy.[5,6] Many studies assess clinical factors which are associated with survival in patients with long bone metastases and some studies incorporate these factors in prediction tools.[7–11] However, we are not aware of studies that consider computed tomography (CT) measurements of body compositions as predictors.

Patients with long bone metastases routinely undergo CTs for staging, assessment of treatment response, or surveillance. These CTs are readily available for analysis and body composition measures could potentially serve as imaging biomarkers to predict outcome in this population without additional risk. Recent studies have proposed CT body composition measurements of muscle and fat depots as biomarkers for survival in patients with and without malignant disease.[12–16]

This study assesses whether body composition measurements obtained using abdominal CTs are independently associated with 90-day and 1-year mortality in patients with long bone metastases undergoing surgery.

# METHODS

### Study Design and Setting

This study complied with the Health Insurance Portability and Accountability Act guidelines. Our institutional review board approved a waiver of informed consent for this retrospective study, performed at a tertiary institution between January 1st, 1999 and January 1st, 2017. We adhered to the Strengthening Reporting of Observational Studies in Epidemiology (STROBE) guidelines.[17]

### Participants and Clinical Characteristics

This single institutional retrospective study, performed at an urban tertiary care referral center for orthopaedic oncology, included: (1) patients 18 years of age or older, (2) surgery for long bone metastases (inclusive of lymphoma and multiple myeloma), and (3) availability of abdominal CT within 3 months prior to surgery[18]. Long bones were defined as femur, humerus, tibia, fibula, radius, and ulna. Excluding criteria were (1) metastatic fractures in multiple bones requiring surgery, (2) revision procedures, (3) surgery other than intramedullary nailing, dynamic hip screw, plate-screw fixation, endoprosthetic reconstruction, or a combination thereof, (4) L4 not included on abdominal CT, and (5) CT not assessable due to metal artifacts. Choice of treatment was decided by mutual agreement between the patient and surgeon, guided by the Mirels score.[19] For patients who underwent multiple CTs within 3 months prior to surgery, only the nearest CT to surgery was included. The first surgery was included if a patient received multiple surgeries meeting the selection criteria. All included CTs were used for determining body composition cross sectional areas (CSA) and only non-contrast CTs for body composition attenuation measurements (Figure 1).

We included 503 patients of which 43% (215/503) had abdominal CTs that were available for body compositions measurements.[20] Of those, 1.4% (3/215) were excluded due to metal artifacts. The remaining 212 CTs were used for determining CSA of subcutaneous abdominal fat (SAT), visceral abdominal fat (VAT) and paraspinal/abdominal muscle. Attenuation measurements were performed on the same three body compositions using non-contrast abdominal CTs (87%; 184/212).

### CT Body Composition Measurements

The CT scanners and body composition measurements were described in detail in our previous study evaluating spinal metastases undergoing surgery.[21] Briefly, measurements were performed at the level of the 4th lumbar vertebra using an in-house automated algorithm, visually inspected and corrected by a single trained observer (CGB) under the supervision of senior fellowship-trained musculoskeletal radiologists (MT, MAB). Body composition measurements included CSA and attenuation in three tissues: VAT, SAT, and paraspinal/abdominal muscle. Muscle CSA was used to determine sarcopenia using total muscle CSA (cm$^2$) divided by the height squared (m$^2$), with cutoff

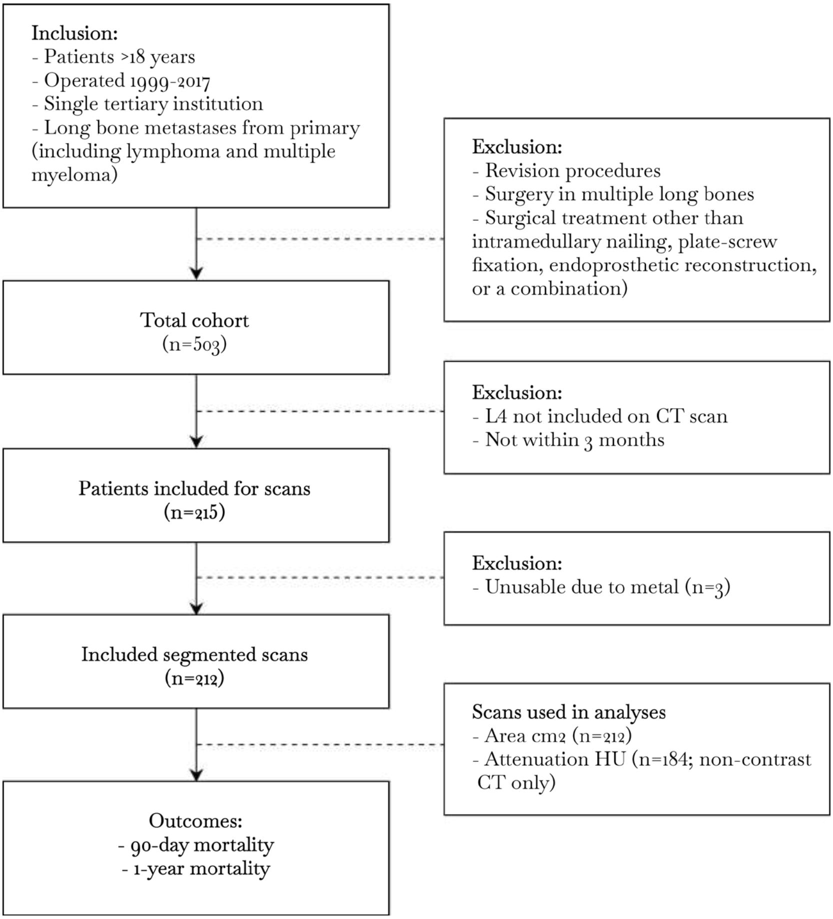

Inclusion:
- Patients >18 years
- Operated 1999-2017
- Single tertiary institution
- Long bone metastases from primary (including lymphoma and multiple myeloma)
Exclusion:
- Revision procedures
- Surgery in multiple long bones
- Surgical treatment other than intramedullary nailing, plate-screw fixation, endoprosthetic reconstruction, or a combination)
Total cohort (n=503)
Exclusion:
- L4 not included on CT scan
- Not within 3 months
Patients included for scans (n=215)
Exclusion:
- Unusable due to metal (n=3)
Included segmented scans (n=212)
Scans used in analyses
- Area cm2 (n=212)
- Attenuation HU (n=184; non-contrast CT only)
Outcomes:
- 90-day mortality
- 1-year mortality

**Figure 1.** Flow diagram depicting the patient selection.

values of <52.4 cm$^2$/m$^2$ for men and <38.5 cm$^2$/m$^2$ for women.[15]

### Outcomes and Explanatory Variables

The outcomes of interest were mortality by any cause after surgery at 90-days and 1-year. Date of death was obtained from medical charts and the Social Security Index.[22] Loss to follow-up in survival was 5.2% (11/212) at 90-days and 8.0% (17/212) at 1-year. Follow-up was verified until May 15th, 2020.

Clinical factors known to be associated with survival[18] were obtained by manual review of medical charts: age, sex, body mass index (BMI, kg/m$^2$), Modified Charlson Comorbidity in addition to metastases[23]; primary tumor categorized as slow, moderate or rapid growth as classified by Katagiri et al.[24], completed pathological fracture, location of long bone metastases (lower or upper extremity), additional metastases to the metastasis operated for, previous systemic therapy, previous local radiotherapy, type of surgical treatment, duration in days of primary tumor diagnosis until metastatic operation, and preoperative albumin level (g/dL) within two weeks of the operation.

### Statistical Analysis

Variables are presented as medians with IQRs for continuous variables and frequencies with percentages for categorical variables. Clinical variables are compared between included and excluded patients using the Mann-Whitney U test for continuous variables and Chi-squared test for categorical variables. Nonparametric testing was used for continuous variables as they were not normally distributed based on inspection of histograms.

Bivariate Cox proportional hazard analysis explored associations between clinical and radiologic variables and the 90-day and 1-year mortality outcomes. We used multivariate Cox proportional hazard analysis including all variables identified in bivariate testing with a P-value of <0.10. The Cox proportionality assumptions and collinearity were tested before performing multivariate analyses. Body composition measurements with a P-value of <0.10 were included separately in the multivariate analyses. Sarcopenia was considered instead of muscle CSA in the multivariate analysis. The Cox results were presented as hazard ratios (HR) with 95% confidence intervals (CI). Kaplan-Meier plots demonstrated the probability of survival for patients with and without sarcopenia. No sample size was calculated since all eligible patients between 1999 and 2017 were included. We applied multiple imputations (n=40) to estimate missing values for BMI in 8 patients (3.8%), and albumin in 5 patients (2.4%). For all analyses, a two-sided P-value of <0.05 was considered significant. All statistical analyses were performed using Stata 15.0 (StataCorp LP, College Station, TX, USA).

# RESULTS

### Study Population

Study patients included 49% (n=103) men and 51% (n=109) women, with a median age of 63 years (interquartile range [IQR], 56-69) and a median BMI of 26 kg/m$^2$ (IQR, 23-30((Table 1) who underwent surgery for long bone metastases. Of the 212 surgeries, 76% (n=162) involved the lower extremities and 24% (n=50) the upper extremities. Systemic therapy was administered prior to surgery in 58% (n=122) and local radiotherapy in 16% (n=33). The four most common primary tumors included lung (22%), renal cell (15%), breast hormone dependent (11%), and multiple myeloma (11%; Appendix 1). The 90-days mortality was 32% (n=64) and 1-year 64% (n=124). Median body composition CSA were for SAT 264 (IQR, 180 – 351) cm$^2$, VAT 134 (IQR, 74 – 816) cm$^2$, and muscle 133 (IQR, 109 – 158) cm$^2$. Median body composition attenuations were for SAT -94 (IQR, -101; -87) Hounsfield unit (HU), VAT -84 (IQR, -90; -69) HU, and muscle 31 (IQR, 23 – 36) HU. Sarcopenia was present in 39% (n=79) patients. The included patients (n=212) differed from the excluded patients (n=291) in the following four characteristics: more comorbidities, more moderate and rapid primary tumor growth, more additional metastases, and a higher 1-year mortality rate (Appendix 2).

We included 503 patients of which 43% (215/503) had abdominal CTs that were available for body compositions measurements.[20] Of those, 1.4% (3/215) were excluded due to metal artifacts. The remaining 212 CTs were used for determining CSA of subcutaneous abdominal fat (SAT), visceral abdominal fat (VAT) and paraspinal/abdominal muscle. Attenuation measurements were performed on the same three body compositions using non-contrast abdominal CTs (87%; 184/212).

### CT Body Composition Measurements

The CT scanners and body composition measurements were described in detail in our previous study evaluating spinal metastases undergoing surgery.[21] Briefly, measurements were performed at the level of the 4th lumbar vertebra using an in-house automated algorithm, visually inspected and corrected by a single trained observer (CGB) under the supervision of senior fellowship-trained musculoskeletal radiologists (MT, MAB). Body composition measurements included CSA and attenuation in three tissues: VAT, SAT, and paraspinal/abdominal muscle. Muscle CSA was used to determine sarcopenia using total muscle CSA ($cm^2$) divided by the height squared ($m^2$), with cutoff values of <52.4 $cm^2/m^2$ for men and <38.5 $cm^2/m^2$ for women.[15]

### Outcomes and Explanatory Variables

The outcomes of interest were mortality by any cause after surgery at 90-days and 1-year. Date of death was obtained from medical charts and the Social Security Index.[22] Loss to follow-up in survival was 5.2% (11/212) at 90-days and 8.0% (17/212) at 1-year. Follow-up was verified until May 15th, 2020.

Clinical factors known to be associated with survival[18] were obtained by manual review of medical charts: age, sex, body mass index (BMI, $kg/m^2$), Modified Charlson Comorbidity in addition to metastases[23]; primary tumor categorized as slow, moderate or rapid growth as classified by Katagiri et al.[24], completed pathological fracture, location of long bone metastases (lower or upper extremity), additional metastases to the metastasis operated for, previous systemic therapy, previous local radiotherapy, type of surgical treatment, duration in days of primary tumor diagnosis until metastatic operation, and preoperative albumin level (g/dL) within two weeks of the operation.

### Statistical Analysis

Variables are presented as medians with IQRs for continuous variables and frequencies with percentages for categorical variables. Clinical variables are compared between included and excluded patients using the Mann-Whitney U test for continuous variables and Chi-squared test for categorical variables. Nonparametric testing was used for continuous variables as they were not normally distributed based on inspection of histograms.

Bivariate Cox proportional hazard analysis explored associations between clinical and radiologic variables and the 90-day and 1-year mortality outcomes. We used multivariate Cox proportional hazard analysis including all variables identified in bivariate testing with a P-value of <0.10. The Cox proportionality assumptions and collinearity were tested before performing multivariate analyses. Body composition measurements with a P-value of <0.10 were included separately in the multivariate analyses. Sarcopenia was considered instead of muscle CSA in the multivariate analysis. The Cox results were presented as hazard ratios (HR) with 95% confidence intervals (CI). Kaplan-Meier plots demonstrated the probability of survival for patients with and without sarcopenia. No sample size was calculated since all eligible patients between 1999 and 2017 were included. We applied multiple imputations (n=40) to estimate missing values for BMI in 8 patients (3.8%), and albumin in 5 patients (2.4%). For all analyses, a two-sided P-value of <0.05 was considered significant. All statistical analyses were performed using Stata 15.0 (StataCorp LP, College Station, TX, USA).

# RESULTS

### Study Population

Study patients included 49% (n=103) men and 51% (n=109) women, with a median age of 63 years (interquartile range [IQR], 56-69) and a median BMI of 26 kg/m$^2$ (IQR, 23-30((Table 1) who underwent surgery for long bone metastases. Of the 212 surgeries, 76% (n=162) involved the lower extremities and 24% (n=50) the upper extremities. Systemic therapy was administered prior to surgery in 58% (n=122) and local radiotherapy in 16% (n=33). The four most common primary tumors included lung (22%), renal cell (15%), breast hormone dependent (11%), and multiple myeloma (11%; Appendix 1). The 90-days mortality was 32% (n=64) and 1-year 64% (n=124). Median body composition CSA were for SAT 264 (IQR, 180 – 351) cm$^2$, VAT 134 (IQR, 74 – 816) cm$^2$, and muscle 133 (IQR, 109 – 158) cm$^2$. Median body composition attenuations were for SAT -94 (IQR, -101; -87) Hounsfield unit (HU), VAT -84 (IQR, -90; -69) HU, and muscle 31 (IQR, 23 – 36) HU. Sarcopenia was present in 39% (n=79) patients. The included patients (n=212) differed from the excluded patients (n=291) in the following four characteristics: more comorbidities, more moderate and rapid primary tumor growth, more additional metastases, and a higher 1-year mortality rate (Appendix 2).

### 90-Day Mortality

Bivariate analysis found that five clinical variables were associated with increased 90-day mortality: lower albumin level, non-white race, comorbidities, rapid primary tumor growth, and previous systemic therapy (all P<0.05). Two body composition measurement were associated with increased 90-day mortality: presence of sarcopenia and lower muscle attenuation (Appendix 3). In multivariate

**Table 1.** Baseline characteristics of patients treated for long bone metastases (n=212).

| *Variables* | *Median (IQR)* |
|---|---|
| Age (years) | 63 (56-69) |
| Body mass index (in kg/m$^2$)[a] | 26 (23-30) |
| Duration primary diagnosis until metastatic operation (months)[a] | 12 (1-41) |
| Pre-operative albumin (g/dL) | 3.7 (3.3-4.1) |
| | *% (n)* |
| Men | 49 (103) |
| Race | |
| White | 92 (195) |
| Non-white | 8 (17) |
| Other Modified Charlson Comorbidity | 69 (147) |
| Primary tumor growth[c] | |
| Slow | 29 (62) |
| Moderate | 29 (61) |
| Rapid | 42 (89) |
| Additional metastases[d] | 87 (185) |
| Tumor location | |
| Upper extremity | 24 (50) |
| Lower extremity | 76 (162) |
| Type of surgery | |
| Intramedullary nail | 45 (96) |
| Endoprosthetic reconstruction | 25 (53) |
| Plate and screw fixation | 25 (53) |
| Dynamic hip screw | 2 (4) |
| Multiple implements | 3 (6) |
| Previous local radiotherapy | 16 (33) |
| Previous systemic therapy | 58 (122) |
| Completed pathological fracture | 56 (118) |
| Mortality[a] | |
| 90-days | 32 (64) |
| 1 year | 64 (124) |
| *Body composition measurements* | *Median (IQR)* |
| Subcutaneous adipose tissue | |
| Area (cm$^2$) | 264 (180 - 351) |
| Attenuation (HU) | -94 (-101; -87) |
| Visceral adipose tissue | |
| Area (cm$^2$) | 134 (74 - 186) |
| Attenuation (HU) | -84 (-90; -69) |
| Muscle | |
| Area (cm$^2$) | 133 (109 - 158) |
| Attenuation (HU) | 31 (23 - 36) |
| | *% (n)* |
| Sarcopenia[e] | 8 (17) |

*IQR=Interquartile range; kg/m$^2$=kilogram per square meter; HU=Hounsfield units*

*a Body mass index was available in 204 patients (96%), albumin in 207 patients (98%), sarcopenia in 205 patients (97%), 90-day mortality in 201 patients (95%), and 1-year mortality in 195 patients (92%).*

*b These values were based on any additional comorbidity on top of the metastatic disease score according to the modified Charlson Comorbidity Index.*

*c Based on histology groupings; slow growth includes hormone dependent breast cancer, hormone dependent prostate cancer malignant lymphoma malignant myeloma, and thyroid cancer; moderate growth includes non-small cell lung cancer with molecularly targeted therapy, hormone independent breast cancer, hormone independent prostate cancer, renal cell carcinoma, sarcoma, other gynecological cancer, and others; and rapid growth includes other lung cancer, colon and rectal cancer, gastric cancer, hepatocellular carcinoma, pancreatic cancer, head and neck cancer, other urological cancer, esophageal cancer, malignant melanoma, gallbladder cancer, cervical cancer, and unknown origin.*

*d Any bone metastasis outside of the lesion treated for.*

*e Sarcopenia cut-off values were <52.4 cm$^2$/m$^2$ (males) and <38.5 cm$^2$/m$^2$ (females).*

analysis after controlling for the five clinical variables, the presence of sarcopenia remained associated with an increased 90-day mortality (HR, 1.87; 95% CI 1.11-3.16; p=0.019; Table 2) but not muscle attenuation (HR, 0.98; 95% CI 0.96-1.00; p=0.079; Appendix 4).

**Table 2.** Multivariable cox proportional hazard analysis for the risk of 90-day death after surgery for long bone metastases using pooled imputed data.

| *Variables* | *Hazard ratio (95% CI)* | *Standard-error* | *P-value* |
|---|---|---|---|
| Albumin | 0.39 (0.26-0.60) | 0.083 | **<0.001** |
| Additional Charlson comorbidity | 1.61 (0.80-3.22) | 0.571 | 0.183 |
| White | 0.46 (0.20-1.06) | 0.196 | 0.068 |
| Primary tumor growth | | | |
| Slow | 0.22 (0.10-0.49) | 0.089 | **<0.001** |
| Moderate | 0.54 (0.30-0.99) | 0.169 | **0.050** |
| Rapid | *Reference value* | | |
| Previous systemic therapy | 1.94 (1.09-3.45) | 0.570 | **0.024** |
| Sarcopenia | 1.87 (1.11-3.16) | 0.499 | **0.019** |

*CI=confidence interval.* Bold *indicates significance (P<0.05).*

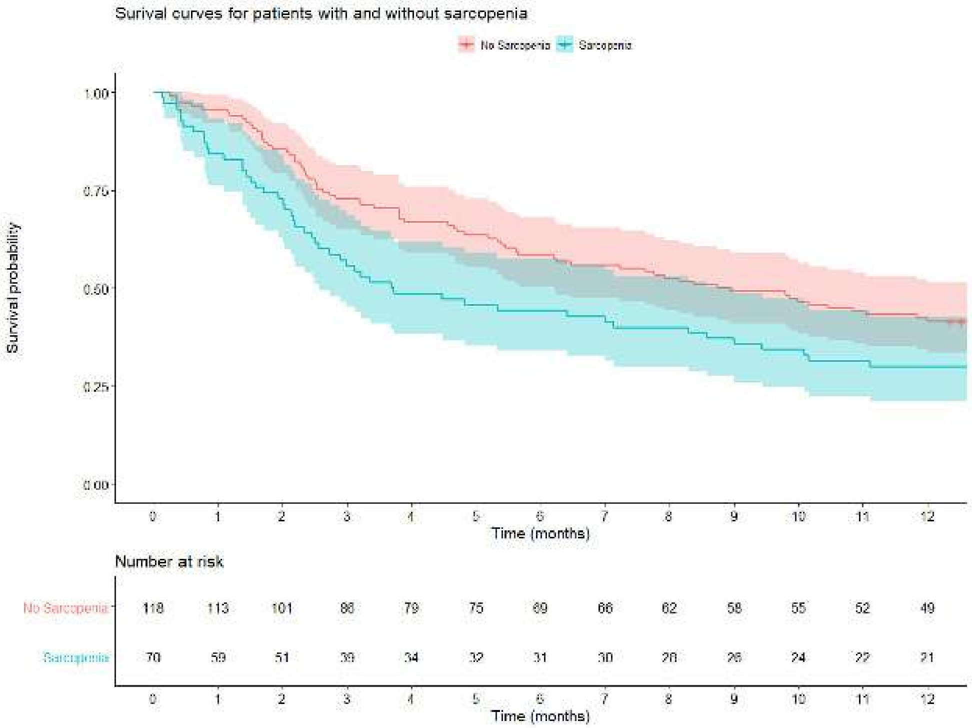
### 1-Year Mortality

Bivariate analysis found that three clinical variables were associated with increased 1-year mortality: lower albumin level, presence of comorbidities and rapid primary tumor growth (all p<0.05). In addition, three clinical variables had a P-value of <0.10: race, additional metastases, and previous systemic therapy. Two body composition measurements were associated with increased 1-year mortality: muscle CSA and presence of sarcopenia, of which the latter was included in multivariate analysis. In multivariate analysis after controlling for six clinical variables, the presence of sarcopenia remained associated with an increased 1-year mortality (HR, 1.50; 95% CI 1.02-2.19; p=0.038; Table 3). The Kaplan-Meier plot illustrated the increased survival probability of patients without sarcopenia (Figure 2).

**Table 3.** Multivariable cox proportional hazard analysis for the risk of 1-year death after surgery for long bone metastases using pooled imputed data.

| *Variables* | *Hazard ratio (95% CI)* | *Standard-error* | *P-value* |
|---|---|---|---|
| Albumin | 0.41 (0.30-0.55) | 0.063 | **<0.001** |
| Additional Charlson comorbidity | 1.54 (0.98-2.42) | 0.354 | 0.060 |
| White | 0.41 (0.20-0.83) | 0.149 | **0.014** |
| Primary tumor growth | | | |
| Slow | 0.20 (0.12-0.33) | 0.053 | **<0.001** |
| Moderate | 0.38 (0.24-0.60) | 0.088 | **<0.001** |
| Rapid | *Reference value* | | |
| Previous systemic therapy | 1.35 (0.89-2.04) | 0.287 | 0.163 |
| Additional metastases | 1.85 (0.89-3.85) | 0.692 | 0.102 |
| Sarcopenia | 1.50 (1.02-2.19) | 0.292 | **0.038** |

*CI=confidence interval.* **Bold** *indicates significance (P<0.05).*



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No Sarcopenia | 118 | 113 | 101 | 86 | 79 | 75 | 69 | 66 | 62 | 58 | 55 | 52 | 49 |
| Sarcopenia | 70 | 59 | 51 | 39 | 34 | 32 | 31 | 30 | 28 | 26 | 24 | 22 | 21 |

**Figure 2.** Kaplan-Meier plot showing the survival probability with 95% confidence intervals for patients with and without sarcopenia.

# DISCUSSION

Survival prognostication is an important element in the surgical decision-making process for patients with long-bone metastases.[9,10,18] Various survival prediction tools have been developed[7,9,18,25,26], but these tools are not optimized and are limited as clinical factors might not be available.[10] Studies concerning patients with extremity sarcoma[13], spinal metastases[12,27,28], and non-osseous malignant neoplasms[29–32] have identified body composition measurements derived from opportunistic CTs as predictor of survival. Our study demonstrates that the presence of sarcopenia is associated with both 90-day and 1-year mortality. Muscle attenuation, SAT and VAT (both attenuation and CSA) were not associated with mortality. To our knowledge, this is the first study assessing CT body composition measurements as predictors of survival in patients surgically treated for long bone metastases while controlling for multiple clinical variables.

This study has several limitations. First, only patients who had undergone a routinely performed CT scan which included the L4 vertebrae were included for analysis. This resulted in the exclusion of 58% of patients (291 of 503) surgically treated for long bone metastases, which may be a source of potential bias. Upon comparison of the patients with and without available CT, we found that patients in the non-CT group had fewer additional Charlson comorbidities, less additional metastases, more primary tumor types with slow growth, and lower 1-year mortality. This suggests that patients without available CT generally consisted of healthier patients with less advanced disease. The prevalence of sarcopenia in this group is unknown which may have impacted the results of this study. However, we believe that this issue has minor impact on the results of this study as we have seen a clear link between sarcopenia and mortality in the frailer patient population included for analysis. Second, there were several factors for which we could not control in multivariate analysis, such as the Eastern Cooperative Oncology Group (ECOG) performance status[33], and preoperative quality of life measures. These measures indicate the preoperative ambulatory status, which may be linked to the amount of skeletal muscle in the patient. Future studies should include these measures to reevaluate and validate these findings. Third, we did not perform analyses of other important secondary outcomes for patients with long bone metastases such as postoperative complications, reoperations, and length of hospitalization. Evidence exists in literature concerning non-osseous neoplasms that body composition measurements has predictive value in these secondary outcomes.[29,34] Last, even though metastases from malignant lymphoma and multiple myeloma are known for their better prognosis, we did include these cases as they formed 16% (33 of 212) of the study cohort.

Sarcopenia, or the involuntary loss of skeletal muscle, has been associated with risk for mortality in patients with various primary malignant neoplasms such as pancreatic, gastric, breast, and lung cancer, in addition to patients suffering from metastatic disease.[35–37] The underlying mechanism that links sarcopenia to mortality in patients with malignant disease has not been well defined. Various candidate mechanisms for muscle wasting have been described – ranging from muscle catabolism due to systemic inflammation, to the inhibition of myoblast differentiation caused by an uninhibited release of the negative muscle cell differentiation regulator, myostatin.[38] However, the involuntary muscle loss is most likely attributed to several simultaneously acting molecular pathways.

The outcomes of this study may have potential implications for clinical care and research. First, our results of the association with sarcopenia with mortality could help clinicians and patients in the shared decision-making process. By integrating the automatically collected body composition biomarker into the electronic health record, clinicians gain yet another aid to better determine optimal treatment for the patient. Second, the finding that the presence of sarcopenia is related to poor survival for both time-points, suggests that involuntary weight loss is not a surrogate for skeletal muscle depletion. Brown et al. showed that, despite body weight stability over time, 1 in 8 patients with colorectal cancer developed incident sarcopenia.[39] Other studies have previously suggested that frailty is better indicated by skeletal muscle loss than decreased body weight.[27,40] Third, apart from the outcomes concerning sarcopenia, primary tumor growth and decreased albumin were found to be independent predictors of a higher risk of mortality in this study. Because these two additional predictors have been successfully incorporated in prognostication tools in previous studies[7], adding the presence of sarcopenia as a variable to these tools may strengthen these predictions.

# CONCLUSION

The presence of sarcopenia assessed by CT is predictive of 90-day and 1-year mortality for patients undergoing surgery for long bone metastases, independent of established risk factors. The presence of sarcopenia could serve as novel biomarker to be included in prediction tools. Future studies should investigate the added benefit of sarcopenia and other opportunistic CT body composition measurements to existing prognostic tools. Accurate and reliable survival prediction is crucial to improve shared decision making for patients with long bone metastases that are considering surgical management.

# REFERENCES

1. Galasko C. The anatomy and pathways of skeletal metastases. *Bone metastases*. 1981:49–63.

2. Coleman RE. Clinical features of metastatic bone disease and risk of skeletal morbidity. *Clin Cancer Res.* 2006;12(20 Pt 2):6243s-6249s.

3. Groot OQ, Ogink PT, Janssen SJ, et al. High risk of venous thromboembolism after surgery for long bone metastases: A retrospective study of 682 patients. *Clin Orthop Relat Res.* 2018;476(10).

4. Janssen SJ, Braun Y, Ready JE, et al. Are allogeneic blood transfusions associated with decreased survival after surgery for long-bone metastatic fractures? *Clin Orthop Relat Res.* 2015;473(7):2343–2351.

5. Wedin R. Surgical treatment for pathologic fracture. *Acta Orthop Scand Suppl.* 2001;72(302):1–29.

6. Willeumier JJ, van der Linden YM, van de Sande MAJ, et al. Treatment of pathological fractures of the long bones. *EFORT open Rev.* 2016;1(5):136–145.

7. Thio QCBS, Karhade AV, Bindels BJJ, et al. Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease. *Clin Orthop Relat Res.* 2020;478(2):322–333.

8. Willeumier JJ, van der Linden YM, van der Wal CWPG, et al. An easy-to-use prognostic model for survival estimation for patients with symptomatic long bone metastases. *J Bone Joint Surg Am.* 2018;100(3):196–204.

9. Bauer HC, Wedin R. Survival after surgery for spinal and extremity metastases. Prognostication in 241 patients. *Acta Orthop Scand.* 1995;66(2):143–146.

10. Janssen SJ, van der Heijden AS, van Dijke M, et al. 2015 Marshall Urist Young investigator award: prognostication in patients with long bone metastases: does a boosting algorithm improve survival estimates? *Clin Orthop Relat Res.* 2015;473(10):3112–21.

11. Sørensen MS, Gerds TA, Hindsø K, et al. Prediction of survival after surgery due to skeletal metastases in the extremities. *Bone Joint J.* 2016;98-B(2):271–7.

12. Kapoor ND, Twining PK, Groot OQ, et al. Adipose tissue density on CT as a prognostic factor in patients with cancer: a systematic review. *Acta Oncol.* 2020;59(12):1488–1495.

13. Veld J, Vossen JA, De Amorim Bernstein K, et al. Adipose tissue and muscle attenuation as novel biomarkers predicting mortality in patients with extremity sarcomas. *Eur Radiol.* 2016;26(12):4649–4655.

14. De Amorim Bernstein K, Bos SA, Veld J, et al. Body composition predictors of therapy response in patients with primary extremity soft tissue sarcomas. *Acta Radiol.* 2018;59(4):478–484.

15. Shachar SS, Williams GR, Muss HB, et al. Prognostic value of sarcopenia in adults with solid tumours: A meta-analysis and systematic review. *Eur J Cancer*. 2016;57:58–67.

16. Brown JC, Cespedes Feliciano EM, Caan BJ. The evolution of body composition in oncology-epidemiology, clinical trials, and the future of patient care: facts and numbers. *J Cachexia Sarcqopenia Muscle*. 2018;9(7):1200–1208.

17. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol.* 2008;61(4):344–349.

18. Nathan SS, Healey JH, Mellano D, et al. Survival in patients operated on for pathologic fracture: Implications for end-of-life orthopedic care. *J Clin Oncol.* 2005;23(25):6072–6082.

19. Mirels H. Metastatic disease in long bones. A proposed scoring system for diagnosing impending pathologic

fractures. *Clin Orthop Relat Res.* 1989;(249):256–264.

20. Groot OQ, Ogink PT, Janssen SJ, et al. High risk of venous thromboembolism after surgery for long bone metastases. *Clin Orthop Relat Res.* 2018;476(10):2052–2061.

21. Bongers ME, Groot OQ, Buckless CG, et al. Body composition predictors of mortality in patients with spinal metastases undergoing surgical treatment. *The Spine Journal.* 2021;23:1529-9430.

22. Social Security Administration. Social Security Death Index. 2014.

23. Quan H, Li B, Couris CM, et al. Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol.* 2011;173(6):676–682.

24. Katagiri H, Okada R, Takagi T, et al. New prognostic factors and scoring system for patients with skeletal metastasis. *Cancer Med.* 2014;3(5):1359–1367.

25. Forsberg JA, Eberhardt J, Boland PJ, et al. Estimating survival in patients with operable skeletal metastases: an application of a bayesian belief network. *PLoS One.* 2011;6(5):e19956.

26. Ratasvuori M, Wedin R, Keller J, et al. Insight opinion to surgically treated metastatic bone disease: Scandinavian Sarcoma Group Skeletal Metastasis Registry report of 1195 operated skeletal metastasis. *Surg oncol.* 2013;22(2):132–8.

27. Zakaria HM, Wilkinson BM, Pennington Z, et al. Sarcopenia as a prognostic factor for 90-day and overall mortality in patients undergoing spine surgery for metastatic tumors: a multicenter retrospective cohort study. *Neurosurgery.* 2020;87(5):1025–1036.

28. Pielkenrood BJ, van Urk PR, van der Velden JM, et al. Impact of body fat distribution and sarcopenia on the overall survival in patients with spinal metastases receiving radiotherapy treatment: a prospective cohort study. *Acta Oncol. (Madr).* 2020;59(3):291–297.

29. Hacker UT, Hasenclever D, Linder N, et al. Prognostic role of body composition parameters in gastric/gastroesophageal junction cancer patients from the EXPAND trial. *J Cachexia Sarcopenia Muscle.* 2020;11(1):135–144.

30. Lee JW, Lee SM, Chung YA. Prognostic value of CT attenuation and FDG uptake of adipose tissue in patients with pancreatic adenocarcinoma. *Clin Radiol.* 2018;73(12):1056.1-1056.10.

31. Rier HN, Jager A, Sleijfer S, et al. Changes in body composition and muscle attenuation during taxane-based chemotherapy in patients with metastatic breast cancer. *Breast Cancer Res Treat.* 2018;168(1):95–105.

32. Rollins KE, Tewari N, Ackner A, et al. The impact of sarcopenia and myosteatosis on outcomes of unresectable pancreatic cancer or distal cholangiocarcinoma. *Clin Nutr.* 2016;35(5):1103–9.

33. Oken MM, Creech RH, Tormey DC, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol.* 1982.

34. Yoon SB, Choi MH, Song M, et al. Impact of preoperative body compositions on survival following resection of biliary tract cancer. *J Cachexia Sarcopenia Muscle.* 2019;10(4):794–802.

35. Argilés JM, Busquets S, Stemmler B, et al. Cancer cachexia: understanding the molecular basis. *Nat Rev Cancer.* 2014;14(11):754–762.

36. Runkel M, Diallo TD, Lang SA, et al. The role of visceral obesity, sarcopenia and sarcopenic obesity on surgical outcomes after liver resections for colorectal metastases. *World J Surg.* 2021.

37. da Cunha LP, Silveira MN, Mendes MCS, et al. Sarcopenia as an independent prognostic factor in patients

with metastatic colorectal cancer: A retrospective evaluation. *Clin Nutr ESPEN.* 2019;32:107–112.

38. Armstrong VS, Fitzgerald LW, Bathe OF. Cancer-associated muscle wasting-candidate mechanisms and molecular pathways. *Int J Mol Sci.* 2020;21(23).

39. Brown JC, Caan BJ, Cespedes Feliciano EM, et al. Weight stability masks changes in body composition in colorectal cancer: a retrospective cohort study. *Am J Clin Nutr.* 2021.

40. Ward MAR, Alenazi A, Delisle M, et al. The impact of frailty on acute care general surgery patients: A systematic review. *J Trauma Acute Care Surg.* 2019;86(1):148–154.

# SUPPLEMENTAL MATERIAL TO CHAPTER 10

**Appendix 1.** Origin of primary tumor (n=212)

**Appendix 2.** Comparison between included (n=212) and excluded, non-CT group (n=291) of patients surgically treated for long bone metastases (n=503).

**Appendix 3.** Bivariate cox proportional hazard analysis for the risk of death (90-days and 1-year) in long bone metastases (n=212) after multiple imputation (n=40).

**Appendix 4.** Multivariable cox proportional hazard analysis for the risk of 90-day mortality for muscle attenuation after surgery for long bone metastases using pooled imputed data.

Supplemental material can be consulted online per the website of the journal and/or publisher

CHAPTER

11

# BODY COMPOSITION PREDICTORS OF ADVERSE POSTOPERATIVE EVENTS IN PATIENTS UNDERGOING SURGERY FOR LONG BONE METASTASES

Peter K. Twining, Olivier Q. Groot, Colleen G. Buckless, Neal D. Kapoor, Michiel E.R. Bongers, Stein J. Janssen, Joseph H. Schwab, Martin Torriani, Miriam A. Bredella

*Journal of the American Academy of Orthopaedic Surgeons (Under review)*

# ABSTRACT

### Background

Body composition assessed by opportunistic, preoperative computed tomography (CT) has been recently identified as a predictor of outcome in patients with cancer.

### Objectives

The purpose of this study was to determine whether cross sectional area (CSA) and attenuation of abdominal subcutaneous adipose tissue (SAT), visceral adipose tissue (VAT), and paraspinous and abdominal muscle are predictors of length of hospital stay (LOS), 30-day postoperative complications, and reoperation in patients treated for long bone metastases.

### Design

Retrospective imaging study.

### Methods

A retrospective database of patients who underwent surgery for long bone metastases from 1999 - 2017 was used to identify 212 patients who underwent abdomen CT. CSA and attenuation measurements for SAT, VAT, and muscle were taken at the level of L4 with aid of an in-house segmentation algorithm. Bivariate and multivariate linear and logistic regression models were created to determine associations between all body composition measurements and outcomes while controlling for confounders including primary tumor, metastasis location, and preoperative albumin.

### Results

On multivariate analysis, increased VAT CSA (regression coefficient(r)[95% CI (confidence interval)]; 0.01 [0.01-0.02]; p<0.01) and decreased muscle attenuation (r [95%CI]; -0.07 [-0.14;-0.01]; p=0.04 were associated with increased LOS. In bivariate analysis, increased muscle CSA was associated with increased chance of reoperation (OR [95% CI]; 1.02 [1.01-1.03]; p=0.04). No body composition measurements were associated with postoperative complications within 30 days.

### Conclusion

Body composition measurements assessed by CT, performed for other purposes, predict adverse postoperative outcomes in patients operated for long bone metastases.

# INTRODUCTION

Treatments for neoplastic disease have rapidly improved over the past several decades and many patients are surviving longer, resulting in increased likelihood of bone metastases.[1] In patients with prolonged expected survival, surgical management is often considered for bone metastases to improve quality of life and protect against impending pathologic fractures[1] and surgical treatment of bone metastases has increased over the last decades.[1] Other treatment strategies of bone metastases include radiotherapy or chemotherapy. As surgery is not without complications, risks and benefits of various treatment options must be thoroughly explored. Many prognostic tools, from simple scoring systems to machine learning algorithms, for predicting mortality after surgical management of metastatic bone disease have been developed to aid surgeons in this decision-making process.[2–9] However, it is also important to consider the possible consequences of surgical management such as prolonged hospital stays, post-operative complications, and reoperations.[10,11] There is a paucity of literature on establishing risk factors for these outcome measures.

Assessment of body composition obtained using computed Tomography (CT) performed for other purposes, so called opportunistic CTs, are able to predict outcome in patients with cancer.[12–14] The most common CT body composition measurements are attenuation and cross-sectional area (CSA) of abdominal subcutaneous adipose tissue (SAT), visceral adipose tissue (VAT), and muscle. Recent studies have shown some of these CT body composition measurements were associated with increased length of hospital stay (LOS), re-admission, post-operative complications, and other adverse outcomes in patients with various gastrointestinal malignancies.[15–18] However, the association of these measurements with adverse postoperative events in patients with long bone metastases undergoing surgery remains unexplored.

The purpose of this study was to determine whether cross sectional area (CSA) and attenuation of abdominal subcutaneous adipose tissue (SAT), visceral adipose tissue (VAT), and paraspinous and abdominal muscle obtained from opportunistic CTs are predictors of length of hospital stay (LOS), 30-day postoperative complications, and reoperation in patients operated on for long bone metastases.

# METHODS

### Patients and Study Design

A retrospective database of patients who underwent surgery for long bone metastases at a single tertiary care center from January 1st, 1999 – December 31st, 2017 was used for this study.[2] Inclusion criteria included (1) patients over 18 years of age, (2) surgically treated for long bone metastatic disease (including lymphoma and multiple myeloma)[6] and (3) pre-operative CT scan including L4 within 3 months prior to the operation. Exclusion criteria were (1) multiple metastatic bone tumors requiring surgery, (2) revision surgeries, (3) surgery type other than intramedullary nailing, endoprosthetic reconstruction, plate and screw fixation, dynamic hip screw, or any combination thereof, or (4) CT scan unusable due to poor quality or no inclusion of the fourth lumbar vertebrae (L4) level.

Surgical decision making for prophylactic fixation in these patients was based on shared decision making guided by the Mirels' score.[19] For patients who underwent multiple abdominal CT scans prior to surgery, the scan closest to the date of operation was chosen. Likewise, for patients who underwent multiple operations for long bone metastases, only the first operation was included.

### CT Analysis

Pre-operative CT scans at the L4 level were used for both cross sectional area (CSA) and attenuation measurements of VAT, SAT, and muscle. For CSA measurements, all scans were used. For attenuation measurements, only non-contrast scans were used. The CT devices, protocol, and methods for analyzing the scans are described more extensively in our previous studies.[20,21] Briefly, scans at the L4 level are analyzed by an automated in-house algorithm and adjusted by a trained researcher (CGB) under the supervision of senior fellowship trained musculoskeletal radiologists (MT, MAB). Body composition measurements were VAT (1) CSA and (2) attenuation, SAT (3) CSA and (4) attenuation, paraspinous and abdominal muscle (5) CSA and (6) attenuation. Sarcopenia was defined as total muscle CSA divided by height squared with cutoff values of $<52.4$ cm$^2$/m$^2$ for men and $<38.5$ cm$^2$/m$^2$ for women.[22]

### Outcomes and Explanatory Variables

Outcome variables were (1) LOS (days), (2) postoperative complications within 30 days, and (3) reoperation. Postoperative complications included pneumonia, venous thromboembolism, sepsis, myocardial infarction, wound infection and/or dehiscence, and urinary tract infection.[23] In addition to CT body composition measurements, variables thought to be associated with post-operative complications were collected from the electronic medical records. These included age, sex, body mass index (BMI), duration from primary diagnosis until operation (days), Charlson Comorbidity Index[24], pre-operative albumin (g/dL), race, primary tumor growth category according to Katagiri [5], additional metastases outside the lesion being treated for, location of bony metastasis (upper or lower extremity), type of surgery, previous radiotherapy, previous systemic therapy, and presence of a pathological fracture.

### Statistical Analysis

Bivariate analysis was used to assess associations of explanatory variables with all three outcomes. Linear regression was used for continuous outcomes (LOS) and logistic regression for categorical outcomes (complications within 30 days and reoperations). All clinical variables with a p value less than 0.10 in bivariate analysis were included in multivariate analysis. Collinearity was tested before performing multivariate analyses and BMI was excluded due to high collinearity with the body composition measurements. Body composition measurements with a p<0.10 were included separately in the multivariate analyses. Multiple imputations (n=40) were applied to estimate missing values for BMI in 8 patients (3.8%) and albumin in 5 patients (2.4%). No multiple imputation was performed for the missing attenuation measurements as this was the explanatory variable of interest. No sample size was calculated since all eligible patients between 1999 and 2017 were included. For all analyses, a two-sided P-value of <0.05 was considered significant. All statistical analyses were performed using Stata 15.0 (StataCorp LP, College Station, TX, USA), R version 3.6.3 (The R Foundation, Vienna, Austria) and R Studio version 1.3.887 (RStudio, Boston, MA).

## RESULTS

### Patients and Characteristics

Of the 503 patients identified who underwent surgery for long bone metastases, 212 had adequate CT scans and met our inclusion criteria to be included in the analysis (Figure 1.) All 212 of these scans were adequate to measure tissue CSA. Only CT scans without intravenous contrast (184) were used for attenuation measurements. Baseline patient characteristics are shown in Appendix 1. The median age of study participants was 63 (interquartile range [IQR] 56-69), with 49% being male and 51% female. Seventy-six percent of patients were treated for lower extremity bone metastases and 24% for upper extremity. The most common primary tumor types were lung (22%), renal (15%), and breast (15%) (Appendix 2). Ninety-day mortality was 32%- and 1-year mortality was 64%. The median LOS was 5 days (IQR 5-7), 10% (21) experienced postoperative complications within 30 days and 4.7% (10) had a reoperation.

The included group had a higher Charlson Comorbidity score, higher proportion of primary tumors in the moderate and rapid tumor growth categories, were more likely to have additional metastases besides the surgically treated lesion, and had a longer LOS than the group excluded due to inadequate or absent CT scans (Appendix 3).

### Length of Stay

On bivariate analysis, increased VAT CSA (regression coefficient (r) [95% confidence interval (CI)]=0.01 [0.010.02]), p=0.01) and decreased muscle attenuation (r [95% CI]=-0.08 [-0.15;-0.01], p=0.03) were associated with longer LOS (Appendix 4). Three clinical variables were controlled for in multivariate analysis: preoperative albumin level, primary tumor growth, and metastasis location. On multivariate analysis, increased VAT CSA (r [95% CI]=0.01 [0.01-0.02], p<0.01) and decreased muscle attenuation (r [95%CI]=-0.07 [-0.14;-0.01], p=0.04) were associated with longer LOS after controlling for all three clinical variables (Table 1 and 2).

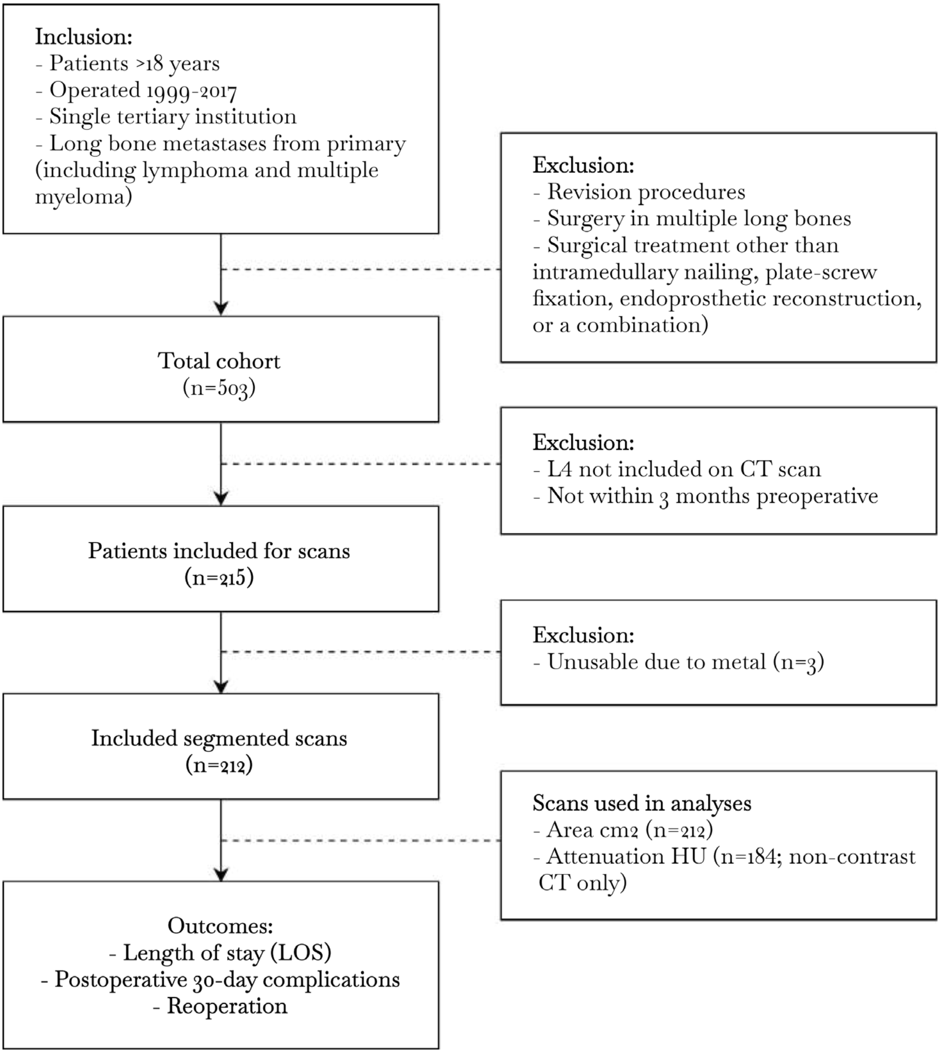

Inclusion:
- Patients >18 years
- Operated 1999-2017
- Single tertiary institution
- Long bone metastases from primary (including lymphoma and multiple myeloma)
Exclusion:
- Revision procedures
- Surgery in multiple long bones
- Surgical treatment other than intramedullary nailing, plate-screw fixation, endoprosthetic reconstruction, or a combination)
Total cohort (n=503)
Exclusion:
- L4 not included on CT scan
- Not within 3 months preoperative
Patients included for scans (n=215)
Exclusion:
- Unusable due to metal (n=3)
Included segmented scans (n=212)
Scans used in analyses
- Area cm2 (n=212)
- Attenuation HU (n=184; non-contrast CT only)
Outcomes:
- Length of stay (LOS)
- Postoperative 30-day complications
- Reoperation

**Figure 1.** Flow diagram depicting the patient selection.

**Table 1.** Multivariable logistic regression analysis for VAT area and length of stay after surgery for long bone metastases using pooled imputed data.

| *Variables* | *Coefficient (95%CI)* | *Standard-error* | *P-value* |
|---|---|---|---|
| Preoperative albumin (g/dL) | -1.95 (-3.04; -0.86) | 0.555 | **<0.01** |
| Primary tumor growth | | | |
| Slow | 0.71 (-0.92; 2.34) | 0.827 | 0.39 |
| Moderate | 2.13 (0.51; 3.75) | 0.823 | **0.01** |
| Rapid | *Reference value* | | |
| Metastasis location | | | |
| Upper extremity | -1.90 (-3.49; -0.31) | 0.808 | **0.02** |
| Lower extremity | *Reference value* | | |
| VAT area (cm$^2$) | 0.01 (0.01; 0.02) | 0.004 | **<0.01** |

*CI=confidence interval; VAT=visceral adipose tissue.* **Bold** *indicates significance (P<0.05).*

**Table 2.** Multivariable logistic regression analysis for muscle attenuation and length of stay after surgery for long bone metastases using pooled imputed data.

| *Variables* | *Coefficient (95%CI)* | *Standard-error* | *P-value* |
|---|---|---|---|
| Preoperative albumin (g/dL) | -1.74 (-3.02; -0.46) | 0.648 | **0.01** |
| Primary tumor growth | | | |
| Slow | 0.44 (-1.43; 2.31) | 0.949 | 0.65 |
| Moderate | 2.40 (0.55; 4.24) | 0.935 | **0.01** |
| Rapid | *Reference value* | | |
| Metastasis location | | | |
| Upper extremity | -1.47 (-3.32; 0.38) | 0.935 | 0.12 |
| Lower extremity | *Reference value* | | |
| Muscle attenuation (HU) | -0.07 (-0.14; -0.01) | 0.034 | **0.04** |

*CI=confidence interval; HU=Hounsfield Units.* **Bold** *indicates significance (P<0.05).*

### Postoperative Complications within 30 Days

On bivariate and multivariate analysis, no explanatory variables or body composition measurements were associated with postoperative complications withing 30 days (Appendix 4).

### Reoperation

On bivariate analysis, increased muscle CSA was associated with increased chance of reoperation (Odds Ratio (OR) [95% CI]; 0.02 [0.01-0.03]; p=0.04, Appendix 4). No controlling for confounders were performed in multivariate analysis as no clinical variables had p values below 0.10. No other body composition measurements were associated with reoperation (p>0.05). Patients that underwent a reoperation had a longer postoperative follow up as compared with patients that did not undergo a reoperation (mean follow up in months: reoperation 20 months versus no reoperation 14 months).

# DISCUSSION

Assessment of body composition measurements on readily available pre-operative CT scans in patients with cancer could provide prognostic information for survival and adverse post-operative outcomes. Body composition measurements from these opportunistic CT scans have been shown to be associated with adverse postoperative outcomes in various patient populations.[15–18] However, to the best of our knowledge, this study is the first study to explore the effects of area and attenuation of SAT, VAT, and muscle on LOS, postoperative complications, and reoperations in patients operated for long bone metastases. Our study shows that (1) increased VAT area and decreased muscle attenuation are associated with longer LOS while controlling for several covariates and (2) increased muscle area is associated with increased chances of reoperation in patients surgically treated for long bone metastases. No association was found between body composition measurements and postoperative complications. This work expands on the growing body of literature that body composition assessed by opportunistic, pre-operative CT may be useful for prognostication in patients with metastatic disease.

This study has several limitations. First, this was a retrospective study and should be interpreted in the appropriate context. To strengthen our retrospective design, CT measurements were made by a researcher blinded to outcomes to mitigate any potential observer bias. Second, despite the large cohort size, during our patient selection process, over 50% of our originally identified cohort was excluded from the analysis due to inadequate on unavailable CT scans. A baseline comparison between the included and excluded groups showed the included group had a higher Charlson comorbidity score, a higher proportion of rapidly growing tumor types, were more likely to have additional metastatic lesions, and a longer LOS. These differences suggest the included group suffered from more advanced disease and comorbidities than the excluded group. It is reasonable that the more fragile group would be more likely to have pre-operative CT scans for cancer staging and surveillance. It is also possible that the results found in this study would not extrapolate to the excluded group. Future studies across various populations would be required to assess the generalizability of these findings. These studies should prospectively include CT- defined body composition measurements to evaluate the prognostic value for these variables' adverse postoperative outcomes. Additionally, improving or maintaining quality of life is recognized as an important outcome to prioritize when evaluating a patient for surgical management of long bone metastases and must be considered alongside survival benefits and risk of complications. Despite these limitations, our large cohort size, in addition with controlling for several known confounding variables lends additional validity to the study.

Our findings that increased VAT area is associated with extended LOS is consistent with previous studies on different disease types in several patient populations.[15,16,18,25,26] Two recent studies of 139 and 110 patients have shown increased VAT area was associated with increased postoperative complications in patients who underwent surgery for gastric or colorectal cancer.[18,26] While our study did not show a relationship between VAT area and postoperative complications, both increased LOS and complications are adverse outcomes that likely result from poor overall health. A study of 2,100 patients, increased VAT area was shown to be associated with increased risk for readmission after surgery for colorectal cancer.[16] In addition, patients operated on for diverticular disease showed an association between increased VAT area and increased postoperative complications.[15] It has been proposed that the adverse effects of visceral adiposity on outcome may be due to its effect on cardiometabolic risk factors including higher incidence of hypertension, diabetes, and metabolic syndrome.[27,28] In patients with colorectal cancer, VAT has been shown to be superior to BMI in predicting the presence of cardiometabolic comorbidities.[29] The consistency of the association of VAT with adverse effects across all these studies of different disease types and surgical locations supports increased VAT area as a marker of poor overall health status.

We found that decreased muscle attenuation was associated with longer LOS. This is consistent with other studies in several different patient populations showing poor outcomes associated with decreased muscle attenuation.[13,30,31] In a study of 805 patients following surgery for colorectal cancer, decreased muscle attenuation was associated with longer LOS.[32] Decreased muscle attenuation is reflective of intramuscular fat deposition, known as myosteatosis, which has been shown to be associated with cancer cachexia.[33,34] In addition, myosteatosis, both in isolation and when combined with visceral obesity, has been shown to be associated with longer LOS in an international cohort of 2,100 patients following surgery for colorectal cancer.[16] Future studies could explore how different combinations of body composition factors may contribute to outcomes. Changes in composition of certain body tissues likely do not occur in isolation, as there has been shown to be a complex cross-talk relationship between adipose tissue and muscle in patients with cancer cachexia.[13,35,36]

Our finding that increased muscle area was associated with increased reoperations is an unexpected finding. A systematic review on patients undergoing abdominal surgery found that low, not high, muscle area was a risk factor for post-surgical adverse events.[37] Another systematic review found sarcopenia, defined as low muscle area, was associated with increased mortality and postoperative complications in surgical oncology patients.[38] This finding appears paradoxical, in that increased muscle area would be associated with reoperation as increased muscle mass is generally present in patients with better overall health status. However, it may be that increased reoperations are a consequence of prolonged survival in this group, as patients with metastatic disease likely have a higher chance of additional operations if they are living longer. In fact, we found that sarcopenia, defined by low abdominal muscle area, was associated with increased 90-day and one-year mortality, lending validity to this theory.[20] To further support this, the mean followup in the group that underwent reoperations was 20 months, versus 14 months in the group that did not undergo reoperations.

Several prognostic models have been developed to assess survival in patients with metastatic bone disease.[2,39,40] However, survival is only one piece of the puzzle, and less emphasis has been given to predicting other adverse postoperative outcomes. To develop scoring systems and algorithms, easily identifiable and interpretable variables associated with those outcomes must be identified. Patients with metastatic bone lesions generally already have preoperative CT scans available, so these so-called opportunistic CT scans can be used to assess body composition measurements which can be incorporated into future prediction models. Future prediction models should consider multiple aspects including survival, risks of adverse outcomes such as complications, length of stay, and reoperations, and potential quality of life benefits to provide surgeons and patients with robust information on which to guide their clinical decisions. We believe that the CT defined body compositions measures presented in this study will be a helpful predictive tool in this prognostication process.

## CONCLUSION

Body composition measurements, assessed by CT performed for other purposes, predict adverse postoperative outcomes in patients operated for long bone metastases. These measures could be incorporated into existing prognostic models to aid physicians and patients in clinical decision making.

# REFERENCES

1. Manabe J, Kawaguchi N, Matsumoto S, et al. Surgical treatment of bone metastasis: indications and outcomes. *Rev Artic Int J Clin Oncol* 2005;10:103–111.

2. Thio QCBS, Karhade AV, Bindels BJJ, et al. Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease. *Clin Orthop Relat Res* 2020;478:322–333.

3. Forsberg JA, Eberhardt J, Boland PJ, et al. Estimating survival in patients with operable skeletal metastases: An application of a Bayesian belief network. *PLoS One* 2011;6(5):E19956

4. Janssen, SJ, Van Der Heijden AS, Van Dijke M, et al. 2015 Marshall Urist Young investigator award: prognostication in patients with long bone metastases: does a boosting algorithm improve survival estimates? *Clin Orthop Relat Res* 2015;473:3112-31221

5. Katagiri H, Okada R, Takagi T, et al. New prognostic factors and scoring system for patients with skeletal metastasis. *Cancer Med* 2014;3:1359–1367.

6. Nathan SS, Healey JH, Mellano D, et al. Survival in patients operated on for pathologic fracture: Implications for end-of-life orthopedic care. *J Clin Oncol.* 2005;23:6072–6082.

7. Willeumier JJ, van der Linden YM, van der Wal CWPG, et al. An easy-to-use prognostic model for survival estimation for patients with symptomatic long bone metastases. *J Bone Jt Surg - Am Vol* 2018;100:196–204.

8. Groot OQ, Bindels BJJ, Ogink PT, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop* 2021;1–9

9. Groot OQ, Ogink PT, Lans A, et al. Machine learning prediction models in orthopedic surgery: A systematic review in transparent reporting. *J Orthop Res* 2021;18

10. Groot OQ, Ogink PT, Janssen SJ, et al. High risk of venous thromboembolism after surgery for long bone metastases: A retrospective study of 682 patients. *Clin Orthop Relat Res* 2018;476:2052–2061.

11. Janssen SJ, Teunis T, Hornicek FJ, et al. Outcome after fixation of metastatic proximal femoral fractures: A systematic review of 40 studies. *J Surg Oncol* 2016;114:507–519.

12. Kapoor ND, Twining PK, Groot OQ, et al. Adipose tissue density on CT as a prognostic factor in patients with cancer: a systematic review. *Acta Oncol (Madr)* 2020;59:1488–1495.

13. Veld J, Vossen JA, De Amorim Bernstein K, et al. Adipose tissue and muscle attenuation as novel biomarkers predicting mortality in patients with extremity sarcomas. *Eur Radiol* 2016;26:4649–4655.

14. De Amorim Bernstein K, Bos SA, Veld J, et al. Body composition predictors of therapy response in patients with primary extremity soft tissue sarcomas. *Acta Radiol* 2018;59:478–484.

15. Tappouni R, Mathew P, Connelly TM, et al. Measurement of visceral fat on preoperative computed tomography predicts complications after sigmoid colectomy for diverticular disease. *Am J Surg* 2015;210:285–290.

16. Martin L, Hopkins J, Malietzis G, et al. Assessment of computed tomography (CT)-defined muscle and adipose tissue features in relation to short-term outcomes after elective surgery for colorectal cancer: a multicenter approach. *Ann Surg Oncol* 2018;25:2669–2680.

17. Hamaguchi Y, Kaido T, Okumura S, et al. Muscle steatosis is an independent predictor of postoperative complications in patients with hepatocellular carcinoma. *World J Surg* 2016;40:1959–1968.

18. Tsukada K, Miyazaki T, Kato H, et al. Body fat accumulation and postoperative complications after abdominal

surgery. *Am Surg* 2004;70:347–351.

19. Mirels, H: Metastatic disease in long bones: A proposed scoring system for diagnosing impending pathologic fractures. *Clin Orthop Relat Res*. 1989;(249);256-264

20. Groot OQ, Bongers MER, Buckless CG, et al. Can Body Composition Measures on Computed Tomography Predict Mortality in Patients with Long Bone Metastases Undergoing Surgery? *Under Rev J Surg Oncol*

21. Bongers MER, Groot OQG, Buckless CG, et al. Body composition predictors of mortality in patients with spinal metastases undergoing surgical treatment. *The Spine Journal.* 2021;23:1529-9430

22. Shachar SS, Williams GR, Muss HB, et al. Prognostic value of sarcopenia in adults with solid tumours: A meta-analysis and systematic review. *Eur J Cancer* 2016;57:58–67.

23. Janssen SJ, Kortlever JTP, Ready JE, et al. Complications after surgical management of proximal femoral metastasis: A retrospective study of 417 patients. *J Am Acad Orthop Surg* 2016;24:483–494.

24. Quan H, Li B, Couris CM, et al. Updating and validating the Charlson Comorbidity Index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol* 2011;173:676–682.

25. Okumura, S, Kaido T, Hamaguchi Y, et al. Visceral adiposity and sarcopenic visceral obesity are associated with poor prognosis after resection of pancreatic cancer. *Ann Surg Oncol* 2017;24:3732–3740.

26. Ozoya OO, Siegel EM, Srikumar T, et al. Quantitative assessment of visceral obesity and postoperative colon cancer outcomes. *J Gastrointest Surg* 2017;21:534–542.

27. Liu J, Fox CS, Hickson DMA, et al. Impact of abdominal visceral and subcutaneous adipose tissue on cardiometabolic risk factors: The Jackson Heart Study. *J Clin Endocrinol Metab* 2010;95:5419–5426.

28. Preis SR, Massaro JM, Robins SJ, et al. Abdominal subcutaneous and visceral adipose tissue and insulin resistance in the framingham heart study. *Obesity* 2010;18:2191–2198.

29. Balentine CJ, Marshall C, Robinson C, et al. Validating quantitative obesity measurements in colorectal cancer patients. *J Surg Res* 2010;164:18–22.

30. Dijk DPJ, Bakens JAM, Coolsen MME, et al. Low skeletal muscle radiation attenuation and visceral adiposity are associated with overall survival and surgical site infections in patients with pancreatic cancer. *J Cachexia Sarcopenia Muscle* 2017;8:317–326.

31. Cushen SJ, Power DG, Murphy KP, et al. Impact of body composition parameters on clinical outcomes in patients with metastatic castrate-resistant prostate cancer treated with docetaxel. *Clin Nutr ESPEN* 2016;13:e39–e45.

32. Malietzis G, Currie AC, Athanasiou T, et al. Influence of body composition profile on outcomes following colorectal cancer surgery. *Br J Surg* 2016;103:572–580.

33. Fujiwara N, Nakagawa H, Kudo Y, et al. Sarcopenia, intramuscular fat deposition, and visceral adiposity independently predict the outcomes of hepatocellular carcinoma. *J Hepatol* 2015;63:131–140.

34. Stephens NA, Skipworth RJE, MacDonald AJ, et al. Intramyocellular lipid droplets increase with progression of cachexia in cancer patients. *J Cachexia Sarcopenia Muscle* 2011;2:111–117.

35. Fearon KCH, Glass DJ, Guttridge DC. Cancer cachexia: Mediators, signaling, and metabolic pathways. *Cell Metab* 2012;16:153–166.

36. Johns N, Greig C, Fearon KCH. Is tissue cross-talk important in cancer cachexia? *Crit Rev Oncog* 2012;17:263–276.

37. Hasselager R, Gögenur I. **Core muscle size assessed by perioperative abdominal CT scan is related to mortality, postoperative complications, and hospitalization after major abdominal surgery: A systematic review.** *Langenbeck's Arch Surg* 2014;399:287–295.

38. Joglekar S, Nau PN, Mezhir JJ. **The impact of sarcopenia on survival and complications in surgical oncology: A review of the current literature.** *J Surg Oncol* 2015;112:503–509.

39. Karhade AV, Thio QCBS, Ogink PT, et al. **Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation.** *Neurosurgery* 2019;85:E671–E681.

40. Schoenfeld AJ, Ferrone ML, Schwab JH, et al. **Prospective validation of a clinical prediction score for survival in patients with spinal metastases: the New England Spinal Metastasis Score.** *Spine J* 2021;21:28–36.

# SUPPLEMENTAL MATERIAL TO CHAPTER 11

**Appendix 1.** Baseline characteristics of patients treated for long bone metastases (n=212)

**Appendix 2.** Origin of primary tumor (n=212)

**Appendix 3.** Comparison between included (n=212) and excluded, non-CT group (n=291) of patients surgically treated for long bone metastases (n=503).

**Appendix 4.** Bivariate linear regression for hospitalization, and bivariate logistic regression for postoperative complications within 30 days, and reoperations in long bone metastases undergoing surgery (n=212)

Supplemental material can be consulted online per the website of the journal and/or publisher

CHAPTER

# 12

# CT DEFINED BODY COMPOSITION MEASUREMENTS FOR PREDICTION OF ADVERSE POSTOPERATIVE EVENTS IN PATIENTS WITH SPINAL METASTASES

Neal D. Kapoor, Olivier Q. Groot, Colleen G. Buckless, Peter K. Twining, Michiel E.R. Bongers, Stein J. Janssen, Joseph H. Schwab, Martin Torriani, Miriam A. Bredella

*Spine (Under review)*

# ABSTRACT

## Background

Computed tomography (CT) body composition measurements have been proposed as biomarkers for postoperative outcomes in patients with and without cancer. The purpose of this study was to determine the value of body composition measures obtained on opportunistic abdomen CTs to predict length of hospital stay (LOS), 30-day postoperative complications, and reoperation in patients with spinal metastases undergoing surgery.

## Objectives

The purpose of this study was to determine whether cross sectional area (CSA) and attenuation of abdominal subcutaneous adipose tissue (SAT), visceral adipose tissue (VAT), and paraspinous and abdominal muscle are predictors of length of hospital stay (LOS), 30-day postoperative complications, and reoperation in patients treated for spinal metastases.

## Design

Retrospective imaging study.

## Methods

Between 2001 and 2016, 196 patients who had surgery for spinal metastases at a single tertiary center underwent CT of the abdomen within three months prior to surgery. Quantifications of cross-sectional areas (CSA) and CT attenuation in Hounsfield Units (HU) of abdominal subcutaneous adipose tissue (SAT), visceral adipose tissue (VAT), and skeletal muscle was performed on CTs. All CT quantification and attenuation measurements were performed at the level of L4. An in-house deep learning algorithm was used to perform the analyses, under the supervision of a fellowship-trained musculoskeletal radiologist. Bivariate and multivariate analyses determined the associations between body composition and outcomes while controlling for clinical variables.

## Results

The median age was 62 years (interquartile range [IQR]=53-70). The median duration of LOS was 9 days (IQR=6-13). 31% (61) of patients had postoperative complications within 30 days, and 16% (31) underwent reoperation. LOS and reoperations were not associated with any CT body composition measurements. Lower muscle CSA (OR [95% CI]=0.99 [0.98-0.99], p=.047) was associated with increased postoperative complication rate after controlling for albumin, thoracic metastases, body mass index, and preoperative neurology score.

**Conclusion**

Body composition measurements may serve as biomarkers for the prediction of postoperative complications in patients with spinal metastases. Future studies should investigate the use of these body composition measurements in the clinical setting by automating it into the electronic healthcare system.

# INTRODUCTION

Medical treatment for patients with cancer has improved considerably over time. As a result, the life expectancy is increasing, and this has the unintended effect of a rising incidence of metastatic disease.[1] In patients with metastatic disease, the spine is a common location, affected in nearly 30% of cases.[2] Spinal metastases can have devastating symptoms, including severe pain, paralysis, incontinence, and sexual dysfunction. Surgical intervention is often indicated for either spinal cord compression or spinal instability.[3–5] With the expanding treatment regimens for these complex patients, multidisciplinary teams and patients must together weigh the likelihood of improved outcomes, including preservation or improved quality of life, against the potential for postoperative morbidity and complications when contemplating surgical management.[6] Prognostic tools can predict these outcomes and thus aid the decision-making process. However, these tools are often limited by lack of clinical variable availability, thus limiting their utility.[7–13]

In recent years, an increasing number of studies has focused on body composition measurements using computed tomography (CT) to predict outcomes in patients with cancer.[14–16] Patients with cancer routinely undergo CT for staging or surveillance of their cancer, and these CTs can be used to assess body composition without additional costs or radiation exposure. This puts CT body composition measurements in a unique category of both being readily available and potentially useful to predict outcomes. Multiple studies have shown these measurements to be useful in predicting survival in various oncologic populations undergoing surgical treatment.[14–18] Recent studies have also shown CT body composition measurements were associated with increased postoperative complications, length of hospital stay (LOS), readmission, and other adverse outcomes in patients with gastrointestinal malignancies.[19–22] However, the predictive value is unknown of body composition measurements for adverse events in patients with spinal metastases undergoing surgery. Identifying new predictors for adverse events is needed as – unlike survival – there is a paucity of data about predicting adverse events in this patient population.

The purpose of our study was to evaluate the value of body composition measurements using abdomen CTs in patients with spinal metastases undergoing surgery to predict LOS, postoperative complications within 30 days of surgery, and reoperations.

# METHODS

### Study Design and Data Sources

This retrospective study was approved by our institutional review board. A waiver of informed consent was approved by our institutional review board for this retrospective study at the tertiary institution Massachusetts General Hospital between January 1st, 2001, and December 31st, 2016. We adhered to the Strengthening Reporting of Observational Studies in Epidemiology (STROBE) guidelines. This study was funded in part by National Institutes of Health Grant K24DK109940 (M.A.B.), and P30DK040561 (M.T.) The authors declare no conflicts of interest.

### Participants and Clinical Characteristics

All patients 18 years of age or older were included that underwent surgical treatment for spinal metastases and had an abdominal CT-scan 3 months prior to surgery. Exclusion criteria were (1) vertebroplasty or kyphoplasty, (2) revision procedures, (3) L4 not included on CT, and (4) CT unreadable due to metal artifacts. The CT scan closest to date of operation was included for patients who underwent multiple abdominal scans. Similarly, in patients who underwent multiple spine operations meeting the selection criteria, only the first surgery was included.

Ultimately, we included 196 patients with suitable CTs to determine cross sectional area (CSA) of abdominal fat and muscle. Of the 196 CTs, 176 (90%) were non-contrast-only, which were used for attenuation measurements in the same abdominal tissues (Figure 1).

### CT Body Composition Measurements

The details regarding CT devices, protocol, and methods for analyzing the scans are described extensively in our previous study. (cite) In brief, CT scans were used for both CSA and attenuation measurements at the L4 level for visceral abdominal tissue (VAT), subcutaneous abdominal tissue (SAT), and paraspinal/abdominal muscle. An automated in-house algorithm analyzed the CTs and were visually corrected by a trained researcher (CGB) under the supervision of senior fellowship trained musculoskeletal radiologists (MT, MAB). In total, we determined six body composition measurements: (1) VAT CSA and (2) VAT attenuation; (3) SAT CSA and (4) SAT attenuation; and (5) muscle CSA and (6) muscle attenuation.

### Outcomes and Explanatory Variables

Outcome were (1) LOS (days), (2) postoperative complications within 30 days, and (3) reoperations until final follow-up or death. We considered the following postoperative complications within 30 days: venous thromboembolism, pneumonia, myocardial infarction, urinary tract infection, sepsis, wound infection and/or dehiscence. Reoperation was defined as unplanned surgical reintervention

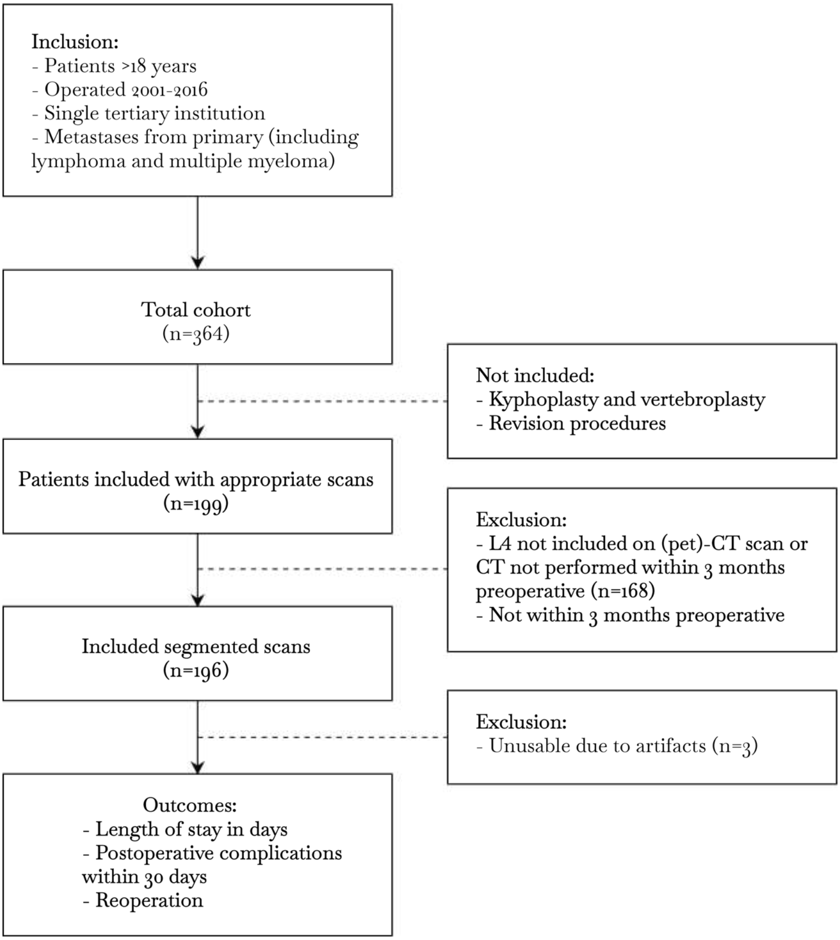

Inclusion:
- Patients >18 years
- Operated 2001-2016
- Single tertiary institution
- Metastases from primary (including lymphoma and multiple myeloma)
Total cohort
(n=364)
Not included:
- Kyphoplasty and vertebroplasty
- Revision procedures
Patients included with appropriate scans
(n=199)
Exclusion:
- L4 not included on (pet)-CT scan or CT not performed within 3 months preoperative (n=168)
- Not within 3 months preoperative
Included segmented scans
(n=196)
Exclusion:
- Unusable due to artifacts (n=3)
Outcomes:
- Length of stay in days
- Postoperative complications within 30 days
- Reoperation

**Figure 1.** Flow diagram depicting the patient selection.

to the initial surgical site. There was no lost to follow-up within 30 days and median follow-up was 9 months (interquartile range [IQR], 3-25 months). Follow-up was verified until May 15th, 2020.

Clinical factors known to be associated with postoperative adverse events or based on expert knowledge were included as explanatory variables by manual chart review: age, sex, body mass index (BMI, kg/m$^2$), any Modified Charlson Comorbidity in addition to metastatic cancer, primary tumor type categorized by Katagiri et al as slow, moderate or rapid growth, tumor location, American Spinal Injury Association (ASIA) impairment scale at time of surgery, additional metastases, spinal pain, previous systemic therapy, region of spinal metastases, completed pathological fracture, type of surgical treatment, duration of primary tumor diagnosis until metastatic operation (days), and preoperative albumin level (g/dL) within two weeks of the operation.

### Patients and Characteristics

Study patients included 123 (63%) men and 63 (47%) women with a median age of 62 years (IQR, 53-70) and BMI of 26 kg/m$^2$ (IQR, 23-30; Table 1). Of the 196 patients, 138 (70%) had additional metastases outside of the lesion treated. The most common primary tumors included renal cell (14%), lung (13%), breast hormone dependent (7.7%), and multiple myeloma (7.1%). Median body composition CSA were for SAT 249 (IQR, 180-320) cm$^2$, VAT 124 (IQR, 75-211) cm$^2$, and muscle 140 (IQR, 116-165) cm$^2$. Median body composition attenuations were for SAT -94 (IQR, -102; -85) HU, VAT -83 (IQR, -90;-60) HU, and muscle 30 (IQR, 23;37) HU. Sarcopenia was present in 42% (76/183) of patients.

**Table 1.** Baseline characteristics of patients surgically treated for spinal metastases (n=196)

| Variables | Spine (n=196) |
|---|---|
| | *Median (IQR)* |
| Age (years) | 62 (53-70) |
| Body mass index (in kg/m$^2$)[a] | 26 (23-30) |
| Duration primary diagnosis untill metastatic operation (days) | 397 (26-1464) |
| Albumin (g/dL) | 3.8 (3.4-4.2) |
| | *% (n)* |
| Male | 63 (123) |
| Additional Modified Charlson Comorbidity[b] | 65 (127) |
| Primary Tumor Growth[c] | |
| Slow | 25 (48) |
| Moderate | 34 (67) |
| Rapid | 41 (81) |
| Additional metastases[d] | 70 (138) |
| Spinal pain | 88 (172) |
| ASIA impairment scale (preoperative)[a] | |
| Neurological deficit (A, B, C, or D) | 45 (88) |
| No neurological deficit (E) | 55 (106) |
| Metastases region | |
| Thoracic | 54 (105) |
| Lumbar | 48 (54) |
| Cervical | 14 (28) |
| Combined | 4.6 (9) |
| Previous local radiotherapy | 33 (64) |
| Previous systemic therapy | 55 (107) |
| Pathological fracture | 54 (106) |
| Number of spine levels undergoing operation | |
| 1 | 47 (93) |
| 2 | 16 (32) |
| 3 or more | 36 (71) |

*Continued on next page*

| | |
|---|---|
| Type of surgery | |
| Vertebrectomy or corpectomy with stabilization | 40 (78) |
| Decompression and stabilization | 39 (76) |
| Decompression | 14 (28) |
| Stabilization | 7.1 (14) |
| Surgical approach | |
| Posterior | 86 (169) |
| Anterior | 11 (22) |
| Combined | 2.6 (5) |
| Two-staged procedure | 1.0 (2) |
| ***Body Composition Measures***[a] | ***Median (IQR) or % (n)*** |
| Non-contrast[e] | 90 (176) |
| Subcutanous adipose tissue | |
| Area ($cm^2$) | 249 (180-320) |
| Attenuation (HU) | -94 (-102; -85) |
| Visceral adipose tissue | |
| Area ($cm^2$) | 124 (75-211) |
| Attenuation (HU) | -83 (-90; -60) |
| Muscle | |
| Area ($cm^2$) | 140 (116-165) |
| Attenuation (HU) | 30 (23-37) |
| Sarcopenia[f] | |
| No | 58 (107) |
| Yes | 42 (76) |
| ***Outcomes*** | ***Median (IQR) or % (n)*** |
| Duration hospitalization in days | 9 (6-13) |
| Postoperative complications within 30 days | 31% (61) |
| Reoperations | 16% (31) |

*IQR=Interquartile range; kg/$m^2$=kilogram per square meter; g/dL=gram per deciliter; ASIA=American Spinal Injury Association; $cm^2$=square centimeters; HU=Hounsfield unit*

*a Body mass index was available in 94% patients (185), ASIA impairment scale in 99% patients (194), SAT area in 80% patients (157), SAT attenuation in 71% patients (140), VAT area in all 100% patients (196), VAT attenuation in 90% patients (176), Muscle area in 80% patients (157), Muscle attenuation in 71% patients (140), and sarcopenia in 77% patients (151).*

*b These values were based on any additional comorbidity on top of the metastatic disease score according to the modified Charlson Comorbidity Index.*

*c Based on histology groupings; slow growth includes hormone dependent breast cancer, hormone dependent prostate cancer malignant lymphoma malignant myeloma, and thyroid cancer; moderate growth includes non-small cell lung cancer with molecularly targeted therapy, hormone independent breast cancer, hormone independent prostate cancer, renal cell carcinoma, sarcoma, other gynecological cancer, and others; and rapid growth includes other lung cancer, colon and rectal cancer, gastric cancer, hepatocellular carcinoma, pancreatic cancer, head and neck cancer, other urological cancer, esophageal cancer, malignant melanoma, gallbladder cancer, cervical cancer, and unknown origin.*

*d Any metastasis outside of the lesion treated for.*

*e Attenuation was measured by only non-contrast CT images.*

*f Sarcopenia cut-off values were <52.4 $cm^2/m^2$ (males) and <38.5 $cm^2/m^2$ (females).*

### Statistical Analysis

Continuous variables are presented as median with IQRs and categorical variables as frequencies with percentages. Linear regression was used to test continuous outcomes (LOS) and with logistic regression for categorical outcomes (complications within 30 days and reoperations). Each separate body composition measurement parameter with p<0.10 was included in multivariate analysis while controlling for all clinical variables that were p<0.10 in bivariate analysis. Collinearity was tested and BMI was excluded from the multivariate analyses because of high collinearity with the body composition measurements. Multiple imputation (n=40) was applied for the following missing variables: BMI in 11 patients (6%) and ASIA score in 2 patients (1%). No multiple imputation was applied for the body composition measurements as this was the variable of interest. No sample size was calculated because all eligible patients in the time period were included. A two-tailed P-value of <0.05 was considered significant. All statistical analyses were performed using R version 3.6.3 (The R Foundation, Vienna, Austria), R studio version 1.3.887 (RStudio, Boston, MA) and Stata 15.0 (StataCorp LP, College Station, TX, USA).

## RESULTS

### Length of Stay

On bivariate analysis, higher SAT attenuation (coefficient [95%CI]=0.06 [-0.01-0.13], p=.06) was not associated with increased LOS but had a P-value of <0.10. The following five clinical variables were included in multivariate analysis: albumin, additional comorbidity, number of spine level

**Table 2.** Multivariable linear regression analysis with SAT attenuation for length of stay (LOS) after surgery for spinal metastases using pooled imputed data.

| *Variables* | *Coefficient (95%CI)* | *Standard-error* | *P-value* |
|---|---|---|---|
| Albumin | -2.00 (-3.74; -0.27) | 0.876 | **0.02** |
| Additional Charlson comorbidity | 0.62 (1.54; 2.77) | 1.089 | 0.57 |
| Number of spine levels undergoing operation | | | |
| 1 | *Reference value* | | |
| 2 | -2.37 (-5.27; 0.53) | 1.464 | 0.11 |
| 3 or more | -3.03 (-5.27; -0.79) | 1.132 | **0.01** |
| Surgical approach | | | |
| Posterior | *Reference value* | | |
| Anterior | -0.91 (-4.00; 2.17) | 1.558 | 0.56 |
| Combined | 1.49 (-5.63; 8.61) | 3.599 | 0.68 |
| Spinal pain | -1.33 (-4.59; 1.94) | 1.650 | 0.42 |
| SAT attenuation | 0.04 (-0.03; 0.11) | 0.034 | 0.24 |

*CI=confidence interval; SAT=subcutaneous adipose tissue.* Bold *indicates significance (P<0.05).*

undergoing operation, surgical approach, and spinal pain (Appendix 1). In multivariate analysis after controlling for the 5 aforementioned clinical variables, SAT attenuation was not associated with LOS (coefficient [95% CI]=0.04 [-0.03-0.11]), p=.24; Table 2).

### Postoperative Complications Within 30 days

On bivariate analysis, higher SAT attenuation (OR [95% CI]=1.02 [1.01-1.05], p=.04) and lower muscle CSA (OR [95% CI]=0.99 [0.98-0.99], p=.03) were associated with increased postoperative complication within 30 days. In addition, lower SAT CSA (OR [95% CI]=0.99 [0.99-1.01], p=.06) had a p<0.10. The following four clinical variables were associated with increased postoperative complications: lower BMI, lower albumin, normal ASIA score, and thoracic metastases. Additionally, previous systemic therapy had a p<0.10. BMI was not included in multivariate analysis due to high collinearity with the body composition measurements. On multivariate analysis, lower muscle CSA (OR [95% CI]=0.99 [0.98-0.99], p=.047) remained associated with increased postoperative complication rate (Table 3). SAT attenuation and SAT CSA were not associated with 30-day postoperative complications while controlling for the four clinical variables (Appendix 2 and 3).

### Reoperations

On bivariate analysis, 3 or more spinal levels undergoing operation (OR [95% CI]=0.36 [0.14-0.95], p=.04) was associated with decreased reoperation rate. No body composition measurements were associated with reoperation (all P>0.10).

**Table 3.** Multivariable logistic regression analysis with Muscle area for 30-day postoperative complications after surgery for spinal metastases using pooled imputed data.

| *Variables* | *Odds ratio (95%CI)* | *Standard-error* | *P-value* |
|---|---|---|---|
| Albumin | 0.42 (0.21; 0.82) | 0.143 | **0.01** |
| ASIA impairment scale (preoperative) | | | |
| Neurological deficit (A, B, C, or D) | *Reference value* | | |
| No neurological deficit (E) | 0.65 (0.30; 1.41) | 0.258 | 0.28 |
| Metastases region | | | |
| Thoracic | *Reference value* | | |
| Lumbar | 0.88 (0.39; 1.98) | 0.363 | 0.76 |
| Cervical | 0.12 (0.02; 0.55) | 0.093 | **0.01** |
| Combined | 0.11 (0.01; 1.17) | 0.133 | 0.07 |
| Previous systemic therapy | 1.27 (0.58; 2.78) | 0.508 | 0.55 |
| Muscle area (cm$^2$) | 0.99 (0.98; 0.99) | 0.006 | **0.047** |

*CI=confidence interval.* Bold *indicates significance (P<0.05).*

# DISCUSSION

The use of CT body composition measurements in prognostication in patients with malignancies who are undergoing surgery is becoming more widespread, and many tools exist to predict survival.[8–11,23] However, there is much less information available on postoperative adverse events, which are essential to the shared decision making between surgeons and potential surgical candidates. To our knowledge, this study is the first of its kind in assessing LOS, postoperative complications within 30 days of surgery, and reoperations in patients with spinal metastases undergoing surgical treatment. We demonstrated on multivariate analysis that lower muscle CSA was associated with increased postoperative complications within 30 days. The current study may serve as a pilot study to demonstrate the value of body composition measurements, which in conjunction with clinical factors, can be used to better predict outcomes after surgery.

This study has several limitations. First, as a retrospective study, there are inherent shortcomings associated with this design. We attempted to mitigate this by blinding outcomes during the CT measurement process. Also, CT measurements are not affected by timing, and we had adequate follow-up for each of our outcomes. Second, surgery is not usually the initial treatment for spine metastases, instead, it usually is utilized when there are complications. Spinal cord compression and impending or unstable pathological fractures are often urgent indications for surgery, and spending time predicting postoperative outcomes may not be useful or appropriate. However, creating the best tools possible to aid physicians and patients in shared decision-making process will be useful to prevent the postoperative morbidity and mortality associated with surgery. Additionally, improving or maintaining quality of life is recognized as an important outcome to prioritize when evaluating a patient for surgical management of spinal metastases. Third, despite this being the largest single-institution cohort of its kind, 46% of the originally identified cohort was excluded from the analysis due to inadequate on unavailable CT scans. A baseline comparison between the included and excluded groups showed the included group had more patients with additional Charlson comorbidities, additional metastases and with pathological fractures (Appendix 4). These differences suggest the included group had more advanced disease and comorbidities than the excluded group. The more fragile group may be more likely to have pre-operative CT scans for cancer staging and surveillance and patients with a history of pathologic fractures may have been diagnosed with imaging modalities other than CT. It is possible that the results found in this population with more advanced disease would not extrapolate to the excluded group. Future, prospective studies with CT body composition measurements across various populations would be required to assess the generalizability of these findings. Last, the impact of radiation therapy on body composition measurements of the tissues surrounding the L4 vertebrae is unknown. However, additional analyses demonstrated that the body composition measurements between lumbar and non-lumbar metastases were not significantly different (data not shown). Despite these limitations, our large cohort size, the number of six different body composition measurements we were able to study, and that we controlled for several known confounding variables lends additional validity to the study. Another important strength was the ability to use an automated in-house algorithm to measure the tissues on the CT scans within seconds, which required minimal correction by clinicians.

The association between lower muscle CSA and greater postoperative complications within 30 days on is consistent with similar oncology studies. Hasselager et al. demonstrate that in patient undergoing abdominal surgery, patients with abdominal and genitourinary malignancies and in very elderly patients undergoing emergency surgery.[24–26] Many studies have also linked low muscle area to decreased survival in a multitude of cancer and non-cancer populations, included patients with extremity and spine metastases.[11,13,18,24,26–28] Our study supports these findings by including the largest cohort to date with various primary tumors, controlling for multiple clinical cofounders, and establishing a convenient and reliable method of collecting the body composition measurements using our in-house algorithm.

Body composition changes seen in patients with metastatic disease may be a result of cachexia, a process of systemic tissue-wasting where quality and quantity of muscle tissue is affected.[29] Cancer causes a hypermetabolic state caused by a mix of tumor metabolism and systemic inflammation that alters the homeostasis of the body, this combined with cancer-related fatigue, anorexia and limited functional status leads to a depletion of skeletal muscle.[25,30] This may be mediated by the inflammatory cytokines tumor necrosis factor-alpha and IL-6, which exert a catabolic effect on muscle by stimulating protein loss in muscle cells.[31,32] The catabolic effect can lead to greater skeletal muscle loss than in other tissues. A better understanding of the role of these inflammatory cytokines and the molecular mechanism behind disproportionate skeletal muscle loss in elderly, surgical, and oncologic populations may help us not only understand the relation of this finding to outcome, but also may point us toward other body composition measurements that can further help us improve our prognostic tools.

The availability of preoperative CT scans in cancer patients and automated algorithms that can quickly and reliably collect body composition analyses make CT based body composition measurements an attractive addition to enhance prognostication tools. Prognostication tools that consider survival and complication factors using both clinical and body composition data may give clinicians the most robust information to work with patients and enhance shared decision making. Future efforts should determine the additional value of these tools in the clinical setting. A robust tool that's proven to be beneficial in the clinical setting could then be added to existing EMR software, automatically inputting the clinical and algorithm obtained body composition data into the model. The model can generate the likelihood of survival and different postoperative complications, allowing surgeons and patients to quickly and clearly determine the best approach to the treatment of their metastatic disease.

# CONCLUSION

In patients with spinal metastases undergoing surgery, lower muscle area is independently associated with increased postoperative complications. Body composition measurements such as lower muscle area could serve as novel biomarkers for prediction of postoperative complications, thereby further optimizing the shared decision-making process. Future studies should investigate the use of these body composition measurements in the clinical setting by automating it into the electronic healthcare system.

# REFERENCES

1. Yahanda AT, Buchowski JM, Wegner AM. **Treatment, complications, and outcomes of metastatic disease of the spine: from Patchell to PROMIS.** *Ann Transl Med.* 2019;7(10):216–216.

2. Ortiz JAO. **The incidence of vertebral body metastases.** *Int Orthop.* 1995;19(5):309–311.

3. Bongers MER, Schwab JH. **Modern Technical Concepts in Surgical Metastatic Disease.** In: Singh K, Colman M, eds. *Surgical Spinal Oncology*. Springer, Cham.; 2020.

4. Patchell RA, Tibbs PA, Regine WF, et al. **Direct decompressive surgical resection in the treatment of spinal cord compression caused by metastatic cancer: a randomised trial.** *Lancet*. 2005;20-26;366

5. Nater A, Sahgal A, Fehlings M. **Management - spinal metastases.** *Handb Clin Neurol.* 2018;149:239–255.

6. Laufer I, Rubin DG, Lis E, et al. **The NOMS framework: approach to the treatment of spinal metastatic tumors.** *Oncologist*. 2013;18(6):744–751.

7. Schoenfeld AJ, Le HV, Marjoua Y, et al. **Assessing the utility of a clinical prediction score regarding 30-day morbidity and mortality following metastatic spinal surgery: the New England Spinal Metastasis Score (NESMS).** *Spine J.* 2016;16(4):482–490.

8. Bongers MER, Karhade AV, Villavieja J, et al. **Does the SORG algorithm generalize to a contemporary cohort of patients with spinal metastases on external validation?** *Spine J.* 2020;20(10):1646–1652.

9. Karhade AV, Thio QCBS, Ogink PT, et al. **Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation.** *Neurosurgery*. 2019;1:85(4);671-681

10. Ghori AK, Leonard DA, Schoenfeld AJ, et al. **Modeling 1-year survival after surgery on the metastatic spine.** *Spine J.* 2015;15(11):2345–2350.

11. Bollen L, van der Linden YM, Pondaag W, et al. **Prognostic factors associated with survival in patients with symptomatic spinal bone metastases: a retrospective cohort study of 1,043 patients.** *Neuro. Oncol.* 2014;16(7):991–8.

12. Tokuhashi Y, Matsuzaki H, Oda H, et al. **A revised scoring system for preoperative evaluation of metastatic spine tumor prognosis.** *Spine (Phila. Pa. 1976).* 2005;30(19):2186–2191.

13. Paulino Pereira NR, Janssen SJ, van Dijk E, et al. **Development of a prognostic survival algorithm for patients with metastatic spine disease.** *J Bone Jt Surg.* 2016;98(21):1767–1776.

14. De Amorim Bernstein K, Bos SA, Veld J, et al. **Body composition predictors of therapy response in patients with primary extremity soft tissue sarcomas.** *Acta Radiol.* 2018;59(4):478–484.

15. Veld J, Vossen JA, De Amorim Bernstein K, et al. **Adipose tissue and muscle attenuation as novel biomarkers predicting mortality in patients with extremity sarcomas.** *Eur Radiol.* 2016;26(12):4649–4655.

16. Kapoor ND, Twining PK, Groot OQ, et al. **Adipose tissue density on CT as a prognostic factor in patients with cancer: a systematic review.** *Acta Oncol.* 2020;59(12):1488–1495.

17. Antoun S, Lanoy E, Iacovelli R, et al. **Skeletal muscle density predicts prognosis in patients with metastatic renal cell carcinoma treated with targeted therapies.** *Cancer*. 2013;119(18):3377–84.

18. Pielkenrood BJ, Urk PR van, Velden JM van der, et al. **Impact of body fat distribution and sarcopenia on the overall survival in patients with spinal metastases receiving radiotherapy treatment: a prospective cohort study.** *Acta Oncol.* 2019;59(3):291–297.

19. Tsukada K, Miyazaki T, Kato H, et al. **Body fat accumulation and postoperative complications after abdominal**

surgery. *Am Surg.* 2004;70(4):347–351.

20. Martin L, Hopkins J, Malietzis G, et al. Assessment of computed tomography (CT)-defined muscle and adipose tissue features in relation to short-term outcomes after Eeective surgery for colorectal cancer: a multicenter approach. *Ann Surg Oncol.* 2018;25(9):2669–2680.

21. Tappouni R, Mathew P, Connelly TM, et al. Measurement of visceral fat on preoperative computed tomography predicts complications after sigmoid colectomy for diverticular disease. *Am J Surg.* 2015;210(2):285–290.

22. Hamaguchi, Y, Kaido, T, et al. Muscle Steatosis is an independent predictor of postoperative complications in patients with hepatocellular carcinoma. *World J Surg.* 2016;40(8):1959–1968.

23. Thio QCBS, Karhade AV, Bindels BJJ, et al. Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease. *Clin Orthop Relat Res.* 2020;478(2):322–333.

24. Hasselager R, Gögenur I. Core muscle size assessed by perioperative abdominal CT scan is related to mortality, postoperative complications, and hospitalization after major abdominal surgery: A systematic review. *Langenbeck's Arch Surg.* 2014;399(3):287–295.

25.Vugt JLA, Levolger L, Coelen RJS, et al. The impact of sarcopenia on survival and complications in surgical oncology: A review of the current literature. *J Surg Oncol.* 2015;112(5):503–509.

26. Du Y, Karvellas CJ, Baracos V, et al. Sarcopenia is a predictor of outcomes in very elderly patients undergoing emergency surgery. *Surgery.* 2014;156(3):521–527.

27. Bernstein KDA, Bos SA, Veld J, et al. Body composition predictors of therapy response in patients with primary extremity soft tissue sarcomas. *Acto Radiol.* 2017;59(4):478–484.

28. Veld J, Vossen JA, Bernstein KDA, et al. Adipose tissue and muscle attenuation as novel biomarkers predicting mortality in patients with extremity sarcomas. *Eur Radiol. 2016 2612.* 2016;26(12):4649–4655.

29. Freire PP, Fernandez GJ, Cury SS, et al. The pathway to cancer cachexia: microRNA-regulated networks in muscle wasting based on integrative meta-analysis. *Int J Mol Sci.* 2019;20(8):1962.

30. Roubenoff, R, Hughes, VA. Sarcopenia: current concepts. *J Gerontol A Biol Sci Med Sci.* 2000;55(12).

31. Tisdale MJ. Wasting in Cancer. *J Nutr.* 1999;129(1):243S-246S.

32. Schaap LA, Pluijm SMF, Deeg DJH, et al. Higher inflammatory marker levels in older persons: associations with 5-year change in muscle mass and muscle strength. *Journals Gerontol Ser A.* 2009;64A(11):1183–1189.

# SUPPLEMENTAL MATERIAL TO CHAPTER 12

**Appendix 1.** Bivariate linear regression for hospitalization, and bivariate logistic regression for postoperative complications within 30 days, and reoperations in spinal metastases undergoing surgery (n=196)

**Appendix 2.** Multivariable logistic regression analysis with SAT attenuation for 30-day postoperative complications after surgery for spinal metastases using pooled imputed data.

**Appendix 3.** Multivariable logistic regression analysis with SAT area for 30-day postoperative complications after surgery for spinal metastases using pooled imputed data.

**Appendix 4.** Comparison between included (196) and excluded, non-CT group (n=168) of patients surgically treated for spinal metastases (n=364).

Supplemental material can be consulted online per the website of the journal and/or publisher.

CHAPTER

13

# THE PREOPERATIVE MACHINE LEARNING ALGORITHM FOR EXTREMITY METASTATIC DISEASE CAN PREDICT 90-DAY AND 1-YEAR SURVIVAL: AN EXTERNAL VALIDATION STUDY OF 264 PATIENTS

Mary K. Skalitzky*, Trevor R. Gulbrandsen*, Olivier Q. Groot*, Aditya V. Karhade, Jorrit-Jan Verlaan, Joseph H. Schwab, Benjamin J. Miller

**Joint first authorship*

*Journal of Surgical Oncology 2021*

# ABSTRACT

### Background

The prediction of survival is valuable to optimize treatment of metastatic long-bone disease. The Skeletal Oncology Research Group (SORG) machine-learning (ML) algorithm for 90-day and 1-year survival has been previously developed and internally validated, however remained to be externally validated.

### Objectives

To determine if the SORG ML algorithm can accurately predicts 90-day and 1-year survival in an independent, external patient cohort surgically treated for metastatic long-bone disease.

### Design

External validation of ML prediction model.

### Methods

A retrospective review of 264 patients who underwent surgery for long-bone metastases between 2003-2019 was performed. Variables used in the stochastic gradient boosting SORG algorithm were age, sex, primary tumor type, visceral/brain metastases, systemic therapy, and ten preoperative laboratory values. Model performance was calculated by discrimination, calibration, and overall performance. The most common primary tumors included renal cell (18%; 47/264), lung (16%, 41/264), and multiple myeloma (14%, 37/264). The mortality, defined as death by any cause, was 19% (51/264) within 90-days and 42% (110/264) within 1-year. The current external validation cohort differed with the SORG development cohort in the following variables: slower growth in primary tumor types, less previous systemic therapy, higher albumin, hemoglobin, absolute lymphocyte count, and white blood cell count, as well as lower 90-day and 1-year mortality.

### Results

Despite the baseline differences, the SORG ML algorithms retained good discriminative ability (area under the curve [AUC] 0.83; 95% confidence interval [CI] 0.76-0.88 for 90-day mortality and AUC 0.84; 95% CI 0.79-0.88 for 1-year mortality), calibration, overall performance, and decision curve analysis.

### Conclusion

The previously developed machine learning algorithms demonstrated good performance in the current study, thereby providing external validation. The models were incorporated into an

accessible application (https://sorg-apps.shinyapps.io/extremitymetssurvival/) that may be freely utilized by clinicians in helping predict survival for individual patients and assist in informative decision-making discussion prior to operative management of long bone metastatic lesions. Future studies are required to validate our algorithms in large, prospective multi-international datasets.

# INTRODUCTION

Recent advancements in medical and surgical therapies for malignancies have resulted in a substantial increase in the incidence of bone metastases. Bone metastases play an important role in overall prognosis, as their occurrence have a significant impact on survival rates and on overall function and quality of life.[2-4] While patients with metastatic bone lesions are usually incurable, operative management should be considered to preserve or improve quality of life for the remaining lifespan. The decision on operative versus conservative management is complex and multifactorial, heavily depending on patient specific factors and estimated overall survival.[5] In patients with appendicular bone metastases, knowing their 90-day and 1-year survival thresholds is critical,[6] as invasive procedures can be detrimental in patients who have a decreased chance of survival past 90-days.[5]

The utilization of a current, validated tool that can accurately predict these survival thresholds is of use for both clinical and shared decision-making discussions. However, as treatment modalities continue to change and survival rates improve, current and validated prognostic models are urgently needed. The Skeletal Oncology Research Group (SORG) machine learning (ML) 90-day and 1-year algorithms have been previously developed to predict survival in patients undergoing surgery for metastatic long bone disease.[7] The 90-day and 1-year time point estimation of survival were chosen to represent short- and long-term survival estimates. While these previously developed algorithms demonstrated good performance, they were created within one healthcare system and have yet to be validated utilizing an external patient population.

In this current study, we set out to determine if the SORG ML algorithm can accurately predicts 90-day and 1-year survival in an independent, external patient cohort surgically treated for metastatic long-bone disease.

# METHODS

## Study Design and Setting

The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD),[8] and the Strengthening the Reporting of Observational Studies in Epidemiology[9] (STROBE) guidelines were followed in this external validation study. After Institutional Review Board approval, an electronic medical record (EMR) retrospective review was performed on all consecutive patients who underwent operative management for long-bone metastases at one academic institution in the United States between 2003 and 2019. The same eligibility criteria, and definitions of outcome and predictors were used as the developmental cohort.

The inclusion criteria consisted of (1) skeletally mature patients, aged 18 years or older, at the time of operative management; (2) surgical treatment for impending or pathological fracture due to metastatic lesions involving appendicular skeletal long bones (humerus, radius, ulna, femur, tibia, fibula) between 2003 and 2019; and (3) primary tumor confirmed by pathology analysis. The exclusion criteria were (1) revision procedures; and (2) surgical treatment other than isolated or combined intramedullary nailing, endoprosthetic reconstruction, dynamic hip screw, or plate-screw fixation. In case of subsequent procedures of the same or separate long bone metastases, only the primary/first procedure was included. In general, operative versus conservative management was determined by the treating surgeon and the patient. Factors considered and discussed included primary tumor type, level of disability and pain, and overall estimated survival.

## Outcomes and Explanatory Variables

The primary outcomes investigated were mortality (yes/no) by any cause at 90-day and 1-year intervals. EMR and Social Security Death Index were reviewed to determine patient survival or the date of last follow-up. The date of the last EMR review was July 15th 2020. Loss to follow-up was 1% (3/264) for 90-day mortality and 6% (15/264) for 1-year mortality.

Variables assessed were included based on necessary input for the SORG ML algorithms, previously determined and validated.[7] This SORG model uses a stochastic gradient boosting algorithm which trains multiple decision trees, and, in a stage, wise fashion optimizes each decision tree by "learning" from mistakes from the previous tree. "Stochastic" indicates that this is done in random order and "gradient boosting" to the optimization of ("learned") weights - the model learns from past mistakes and adjusts the weights of the accurately and wrongly classified data until further improvement is not possible.[10] The following predictive variables were obtained preoperatively: age; sex; body mass index (BMI; kg/m$^2$); primary tumor histology (based on groupings by Katagiri et al.[11]); tumor location; multiple bone metastases; visceral metastases (liver and/or lung); brain metastases; previous systemic therapy; and local radiation therapy. Preoperative laboratory values within two weeks of procedure were recorded including: albumin level (g/dL), alkaline phosphatase (IU/L), calcium (mg/dL), creatinine (mg/dL), hemoglobin level (g/dL), lymphocyte absolute count (x $10^3$/uL), neutrophil absolute count (x $10^3$/uL), platelet count (x $10^3$/uL), sodium (mg/dL) levels, and white blood cell count (x $10^3$/ uL). Prediction variables and outcome data were obtained blinded from one another by different extraction datasheets.

### Statistical Analysis

Baseline differences between the validation and development cohort were assessed using the chi square test for categorical variables and Mann-Whitney U test for continuous variables. A two-tailed p value < 0.05 was considered significant.

For each patient, an individual predicted probability was formulated for 90-day and 1-year survival by inputting all 15 predictive variables into the SORG algorithm. Next, the predictive probabilities were analyzed and compared with the actual outcomes at 90-days and 1-year. The performance of the SORG algorithm was assessed by the metrics utilized in the developmental study following the TRIPOD guidelines. The TRIPOD guidelines consist of the following: (1) discrimination (area under the curve (AUC), and F1-score); (2) calibration using a plot, intercept and slope; (3) overall performance using Brier score and null model Brier score; and (4) decision curve analysis.[7, 12] Discrimination was assessed by calculating the AUC and visualized by plotting the receiver operating characteristic curve (whereby an AUC of 0.5 represented chance, indicating no discrimination, whereas an AUC of 1.0 represented perfect discrimination). The Youden index was calculated across different threshold, allowing for the selection of the threshold that maximizes the sum of sensitivity and specificity.[13] F1-score calculates the overall accuracy of the algorithm, which ranges between 0 (total failure) and 1 (perfect algorithm). The F1-score can be interpreted as an equal contribution of both precision and recall. Precision, also known as the positive predictive value, refers to the proportion of the true positives on all positive predictions. A precision of 1 means that there are no false positives. Recall, also known as sensitivity, is the proportion of positives correctly predicted. A recall of 1 corresponds with no incorrect negative predictions. Thus, a F1-score nearing the 1 means that there are low false positives and low false negatives, indicating that the algorithm is correctly predicting the actual mortalities.[14]

Calibration referred to how closely the predicted 90-day and 1-year mortality agreed with the observed outcomes.[15, 16] This was visualized by plotting the predicted probabilities (x-axis) against the observed frequencies (y-axis) of the outcome. In this plot, perfect predictions should lie on the 45° upward line for complete agreement with the outcome, with a slope of 1.0 and an intercept of 0. A slope greater than 1.0 would indicate overfitting and a slope lower than 1.0 would indicate underfitting. For example, a slope of 0.8 indicated that predicted 90-day or 1-year mortality rates were on average too high for patients with high probabilities and too low for patients with low probabilities. The calibration intercept indicated the overall tendency for underestimation (positive values) or overestimation (negative values) of the outcome.[17] For example, a negative calibration intercept would represent an overestimation of the predicted 90-day or 1-year mortality risk compared with the observed proportion. The Brier score is a way to verify the accuracy of the predictions calculated by the algorithm, ranging from 0 (excellent prediction) to 1 (worst prediction).[15, 16, 18] To allow correct interpretation, a comparison needed to be made with the null-model Brier score, which assigned a predicted probability equal to the observed prevalence of 90-day and 1-year mortality in this external validation cohort. A Brier score lower than the null-model Brier score represented greater performance of the SORG algorithm. Decision curve analysis provided a visual comparison of different treatment strategies to establish the net benefit (weighted average of true positives and false positives) across a range of different threshold probabilities. The horizontal "none" line represented the net benefit without any changes in management; no benefit or harm is expected from this strategy. The slanted "all" line represented the net benefit with treatment change across all patients. The future user of the algorithm can establish which threshold is important and decide if the predicted net benefit at that particular threshold is valuable.

The missForest method was utilized for multiple imputation, in the setting of missing data.[19] Missing data included: platelet count 1% (3/264), hemoglobin level 1% (3/264), white blood cell count 1% (3/264), sodium 2% (5/264), creatinine 3% (8/264), calcium 4% (11/264), lymphocyte absolute count 8% (21/264), neutrophil absolute count 8% (21/264), albumin 23% (62/264), and alkaline phosphatase 23% (70/264). None of the variables had more than 30% missing data.

No sample size was calculated since all eligible patients between 2003 and 2019 were included. Statistical software used for data analysis and model validation were Stata 15 (StataCorp LP, College Station, TX, USA) and R version 3.5.1 (The R Foundation, Vienna, Austria).

# RESULTS

Overall, 264 patients including 137 (52%) women and 127 (48%) men were included in this study. The median age of all patients was 64 years (interquartile range [IQR], 54–71; Table 1). The most common primary tumors were renal cell (18%; 47/264), lung (16%; 41/264), and multiple myeloma (14%; 37/264). The mortality was 19% (51/264) within 90-days and 42% (110/264) within 1-year. The current external validation cohort differed with the initial SORG development cohort in the following eight variables: slower growth in primary tumor types, less previous systemic therapy, higher albumin level, higher hemoglobin level, higher lymphocyte absolute count, higher white blood cell count, and lower 90-day and 1-year mortality. These differences demonstrate that this external validation cohort was different in both patient and disease characteristics. There were no differences in eligibility criteria

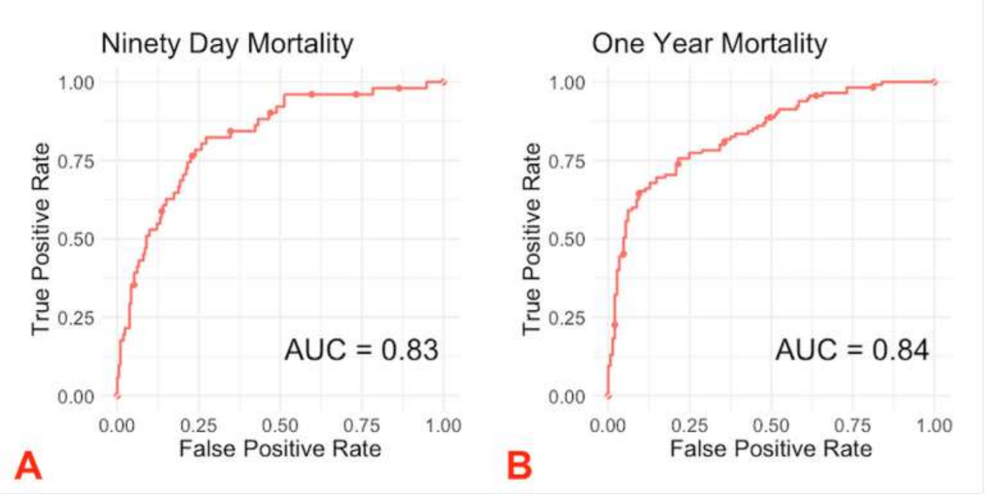


**Figure 1.** Flow diagram depicting the patient selection.

as the same inclusion and exclusion criteria were used in both the developmental and validation cohort, and both hospitals tertiary care centers.

The 90-day SORG ML algorithm achieved an AUC of 0.83 (95% confidence interval [CI] 0.76-0.88; Figure 1A and Table 2) in this external validation cohort. The calibration plot demonstrated excellent calibration from predicted probability 0 to 0.7 (Figure 2A). Per predicted probabilities greater than 0.7, the algorithm overestimated the proportion of patients with 90-day mortality, which is reflected in the overall negative intercept of -0.21 (95% CI -0.58-0.17) and a slope of 0.84 (95% CI 0.59-1.09). The Brier score for the overall algorithm performance was 0.12 (95% CI 0.10-0.15) compared with a higher null-model Brier score of 0.16 indicating greater performance of the SORG algorithm. Decision curve analysis provided greater standardized net benefit at all predicted probabilities compared to default strategies of changing management for all patients or no patients (Figure 2C). In other words, above a high-risk threshold of 0.2 the predictions of the algorithm resulted in a larger net (survival) benefit compared to changing the treatment for all patients or for no patients.

The 1-year SORG ML algorithm achieved an AUC of 0.84 (95% CI 0.79-0.88; Figure 1B). The calibration plot showed good calibration for predicted probability less than 0.3 and greater than 0.9 (Figure 2B). For predicted probabilities between 0.3 and 0.9, the algorithm overestimated the proportion of patients with 1-year mortality, which is reflected in the overall negative intercept of -0.73 (95% CI -0.1.02-0.44) and a slope of 1.08 (95% CI 0.81-1.35). The Brier score for the overall algorithm performance was 0.18 (95% CI 0.16-0.21) compared with a higher null-model Brier score of 0.25 indicating greater performance of the SORG algorithm. Decision curve analysis provided greater standardized net benefit at all predicted probabilities compared to default strategies of changing management for all patients or no patients (Figure 2D).

**Table 1.** Baseline comparison between external validation (n=264) and development population (n=1090)

| Variables | Validation (n=264) | Development (n=1090) | |
|---|---|---|---|
| | % (n) or median (IQR) | | P-value |
| Age | 64 (54-71) | 63 (54-72) | 0.76 |
| Female sex | 52% (137) | 56% (610) | 0.23 |
| Primary tumor type[a] | | | **0.03** |
| Slow growth | 43% (114) | 42% (460) | |
| Moderate growth | 31% (81) | 24% (263) | |
| Rapid growth | 26% (69) | 34% (367) | |
| Visceral metastases | 46% (121) | 45% (487) | 0.74 |
| Brain metastases | 13% (35) | 16% (175) | 0.26 |
| Previous systemic therapy | 55% (144) | 62% (676) | **0.03** |
| Tumor location | | | 0.50 |
| Upper extremity | 25% (67) | 23% (255) | |
| Lower extremity | 75% (197) | 77% (835) | |
| *Preoperative laboratory values*[b] | | | |
| Albumin level (g/dL)[c] | 4 (3-4) | 4 (3-4) | **0.01** |
| Alkaline phosphatase level (IU/L) | 104 (81-137) | 101 (74-146) | 0.49 |
| Calcium (mg/dL) | 9 (9-10) | 9 (9-10) | 0.14 |
| Creatinine (mg/dL) | 0.8 (0.7-1.1) | 0.8 (0.7-1.1) | 0.47 |
| Hemoglobin level (g/dL) | 12 (10-14) | 11 (10-13) | **0.00** |
| Lymphocyte absolute count ($10^3$/uL)[c] | 1 (1-2) | 1 (1-2) | **0.01** |
| Neutrophil absolute count ($10^3$/uL) | 5 (4-8) | 5 (4-8) | 0.91 |
| Platelet count ($10^3$/uL) | 240 (184-308) | 251 (184-332) | 0.24 |
| Sodium (mg/dL) | 138 (136-141) | 138 (136-140) | 0.09 |
| White blood cell count ($10^3$/uL) | 8 (6-11) | 7 (5-10) | **0.01** |
| Mortality[b] | | | |
| 90-day | 19% (51) | 29% (305) | **0.00** |
| 1-year | 42% (110) | 62% (639) | **<0.01** |

*Baseline characteristics were compared using the chi-square test for categorical variables and Mann-Whitney U test for continuous variables.* **Bold** *indicates significance (P<0.05).*

*IQR=interquartile range; g/dL=grams per deciliter; IU/L=international units per liter; kg/m$^2$=kilograms per meter squared; mg/dL=milligrams per deciliter; uL=microliter*

*a Slow growth includes hormone dependent breast cancer, hormone dependent prostate cancer malignant lymphoma malignant myeloma, and thyroid cancer; moderate growth includes non-small cell lung cancer with molecularly targeted therapy, hormone independent breast cancer, hormone independent prostate cancer, renal cell carcinoma, sarcoma, other gynecological cancer, and others; and rapid growth includes other lung cancer, colon and rectal cancer, gastric cancer, hepatocellular carcinoma, pancreatic cancer, head and neck cancer, other urological cancer, esophageal cancer, malignant melanoma, gallbladder cancer, cervical cancer, and unknown origin.*

*b Missing data in validation cohort: albumin 23% (62/264), alkaline phosphatase 27% (70/264), calcium 4% (11/264), creatinine 3% (8/264), hemoglobin 1% (3/264), lymphocyte count 8% (8/264), neutrophil count 21% (21/264), platelet count 1% (3/264), sodium 2% (5/264), white blood cell count 1% (3/264), vital status at 90-days 1% (3/264), and 1-year 6% (15/264).*

*c The validation cohort had a higher albumin level and lymphocyte absolute count than the developmental cohort*

Table 2. Performance of SORG machine learning algorithms for extremity metastasis on external validation (n=264)

| *Performance metric* | *90-day mortality* | *1-year mortality* |
|---|---|---|
| | *Discrimination* | |
| AUC | 0.83 (0.76, 0.88) | 0.84 (0.79, 0.88) |
| F1-score[a] | 0.56 (0.44, 0.67) | 0.72 (0.63, 0.81) |
| | *Calibration* | |
| Intercept | -0.21 (-0.58, 0.17) | -0.73 (-1.02, -0.44) |
| Slope | 0.84 (0.59, 1.09) | 1.08 (0.81, 1.35) |
| | *Overall* performance | |
| Brier score | 0.12 (0.10, 0.15) | 0.18 (0.16, 0.21) |
| Null-model Brier score | 0.16 | 0.25 |

*AUC=area under the receiver operating curve.*
*a probability threshold equal to the Youden index (90-day mortality threshold=0.19, 1-year mortality threshold=0.76)*

An example of the SORG model predicting 1-year survival probability of 32% of an externally validated patient is shown in Figure 3. The variables that favored survival are visualized by the green bars: no brain metastases, alkaline phosphatase level between 82 and 107 IU/L, moderate-growth primary tumor, and platelet count between 193 and 258 x 103/uL. The variables that resulted in an adjustment that increased the probability of mortality are visualized by the red bars: sodium level higher than 136 mg/dL, albumin level between 3.30 and 3.70, and neutrophil-to-lymphocyte ratio between 5.29 and 7.76. The clinical characteristics of each individual patient can be filled in to provide a survival prediction in real time. This model can be accessed at https://sorg-apps.shinyapps.io/extremitymetssurvival/.

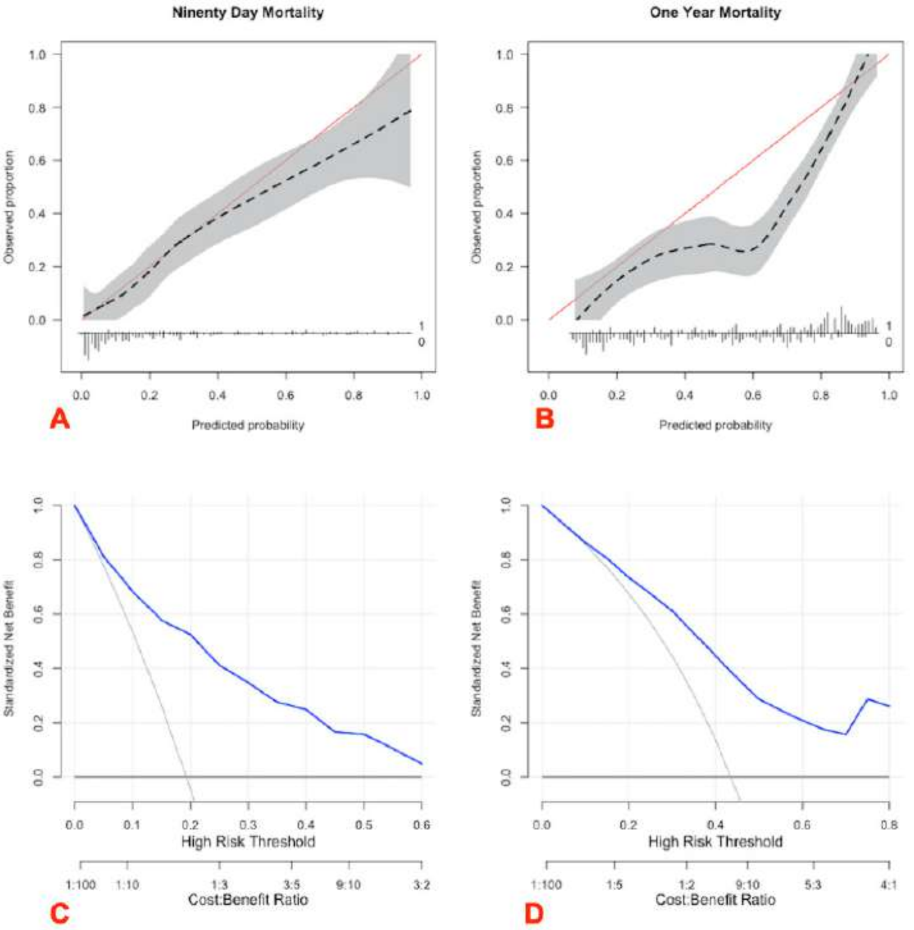
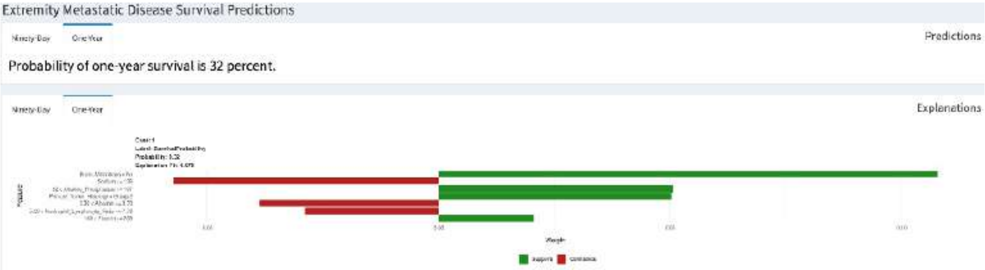


**Figure 2.** Calibration (A-B) and decision curve analysis (C-D) of SORG machine learning algorithm for 90-day and 1-year mortality on external validation, n=264.

# DISCUSSION

In our external validation of the SORG ML algorithms with 264 patients, we found that the algorithm was able to accurately predict 90-day and 1-year mortality in an external patient population. As advances in oncological treatment continue, knowing the overall prognosis is a valuable tool to aid clinicians and patients in decision making in patients with bone metastases. The formerly developed SORG ML algorithm has not been previously externally validated. In accordance with the TRIPOD guidelines, external validation is *"an invaluable and crucial step in the introduction of a new prediction* *model before it should be considered for routine clinical practice"*.[12, 20] Toward this end, the SORG ML algorithms were externally validated in an independent population, taking one step further to implementation in clinical care. Future validation in large, prospective multi-international datasets is warranted to further validate or refute these algorithms.



Extremity Metastatic Disease Survival Predictions

Ninety-Day | One-Year

Predictions

Probability of one-year survival is 32 percent.

Ninety-Day | One-Year

Explanations

**Figure 3.** Web application interface for explanation of variables that supports (green) or contradicts (red) 1-year survival for an individual patient at https://sorg-apps.shinyapps.io/extremitymetssurvival/.

This study has several limitations. First, the patients were retrospectively included from a single institution. Due to the relatively limited volume of treated musculoskeletal lesions of the present institution, the external validation patient cohort was smaller than the initial internal validation. In addition, both the developmental and external validation cohort were from the same nation. Although geographically distinct, the algorithm remains to be tested in a non-American population since treatment, patient and disease characteristics may be different in other countries. For example, tumor biology has shown large variations by race and ethnicity, and access to care and quality of (surgical) cancer treatment might be very distinct else in the world.[27] Future research in large, prospective international datasets are needed to address these concerns. Second, another limitation of the relatively small dataset was that recalibration to improve the model was not possible. Although current (national) registries have sufficient patient data, they lack the required high-quality input variables such as the presence of visceral metastases or preoperative laboratory values. Future multicenter efforts should seek to create registries of patients with both sufficient volume and granular data to ensure reliable interpretation and potential recalibration.[28, 29] Third, there were differences in baseline characteristics between the two patient populations such as a lower proportion of rapidly growing tumors, lower rates of previous systemic therapy, and lower mortality at both 90-days and 1-year. Although the reasons for these differences are unknown as both institutions are tertiary care hospitals from the same nation, it does demonstrate that the algorithm maintains accurate discriminative ability and overall performance across varying patient and tumor populations. Fourth, the relatively long inclusion of the cohort (2003-2019) may have had impact on our results as advancements in oncology have been tremendous in the last 10 years and the cohort is therefore relatively 'old' and may not have benefitted from all these new treatments. However, the developmental cohort consisted of patients from a similar timespan (1999-2017) thereby taking these differences into account. In addition, an additional analysis stratified by year of treatment (prior to 2015 and after 2015) demonstrated no differences in AUCs (data not shown) indicating that the algorithm is generalizable to current patients that are considering treatment. Finally, this tool was developed to predict mortality in patients with long-bone metastases who were managed surgically for (impending) pathologic fractures. The applicability of this algorithm to other treatment options has not yet been studied. Future research would benefit from validating this algorithm with a prospective patient population, as well as study the efficacy of this algorithm with non-operative management. Nevertheless, this study is the first external validation of the SORG ML algorithm and currently among the best performing externally validated prediction tools for 90-day and 1-year mortality in patients with long bone metastasis. To improve surgical decision-making, accurate and reliable externally validated survival tools are required such as this SORG algorithm.

The SORG ML algorithms retained good discrimination and overall performance in this external validation among a contemporary cohort of consecutively treated patients for long bone metastases. The discrepancies in the calibration results, illustrated by the overestimation of both 90-day and 1-year mortality in certain prediction probabilities, between the external validation and developmental cohort may be explained in two ways. First, the external validation cohort differed with the SORG development cohort in the patient and disease characteristics: primary tumor type, previous systemic therapy, albumin level, hemoglobin level, lymphocyte absolute count, and white blood cell count. Second, because of the relatively small sample size, the proposed minimum of 200 events and non-events for both outcome groups were not met with this external validation cohort.[21] This minimum proposed balance of outcome events is required for reliable interpretation of the calibration results and may explain the overestimation as shown by the calibration plots, resulting in an intercept of -0.21 and -0.73, compared with the 0.06 and 0.09 from the developmental dataset, for 90-day and 1-year mortality, respectively. For now, the fact that the SORG models correctly orders patients according to their risks in both 90-day and 1-year mortality, reflected by the good AUCs, but does not provide a fair estimate of that risk, reflected by the overestimation in the calibration curves, calls for further work. Larger studies can consider recalibrating the SORG models to further improve on calibration. This highlights the need for multi-institutional collaboration to achieve large datasets with the required high-quality input variables.

Various survival prognostication tools exist for patients with long bone metastatic disease.[6, 11, 22-25] These non-externally validated models all performed worse to the algorithm externally validated here and were thoroughly discussed (e.g. differences in included predictive factors) upon development and internal validation.[7] To our knowledge, only one prognostication tool has been externally validated for 90-day and 1-year survival.[26] A Bayesian belief network was developed in 189 patients surgically treated for bone metastases to the extremities.[6] On external validation of 815 patients, slightly lower AUCs of 0.79 and 0.76 were achieved for 90-days and 1-year mortality, respectively.[26] Both external validations demonstrated in decision curve analysis that greater standardized net benefit was achieved at all predicted probabilities compared with default strategies of changing management for all patients or no patients. Although the calibration plots showed reasonably well calibrations for prediction of 90-day and 1-year mortality, further essential measures such as the calibration slope and intercept were missing in order to compare performance and assess quality of the external validation.[20] Especially since the calibration results would be even more of interest since their validation cohort consisted of 200 evens and non-events in both outcome groups unlike our sample size. Incomplete presentation of the performance measures emphasizes the need for clear and transparent reporting of model validation following the prescribed TRIPOD guidelines.[8]

## CONCLUSION

The previously developed ML algorithms to predict 90-day and 1-year survival in patients who underwent surgery for long bone metastases demonstrated good discriminative capability and overall performance in this study, providing external validation. These models are incorporated into an accessible application that can be found at https://sorg-apps.shinyapps.io/extremitymetssurvival/. This application may be freely utilized by clinicians in predicting survival for individual patients and may assist in informative decision-making discussion with the patients prior to operative management of long bone metastatic lesions. Future validation in large, prospective multi-international datasets, especially including non-American patients, is warranted to further validate or refute these algorithms.

# REFERENCES

1. Saucedo JM, Vedder NB. Firework-related injuries of the hand. *J Hand Surg Am*. 2015;40(2):383-7.

2. Falk S, Bannister K, Dickenson AH. Cancer pain physiology. *Br J Pain*. 2014;8(4):154-62.

3. Middlemiss T, Laird BJ, Fallon MT. Mechanisms of cancer-induced bone pain. *Clin Oncol (R Coll Radiol)*. 2011;23(6):387-92.

4. Ogura K, Yakoub MA, Christ AB, et al. What are the minimum clinically important differences in SF-36 scores in patients with orthopaedic oncologic conditions? *Clin Orthop Relat Res*. 2020:478;2148-2158

5. Wedin R. Surgical treatment for pathologic fracture. *Acta Orthop Scand Suppl*. 2001;72(302):2p, 1-29.

6. Forsberg JA, Eberhardt J, Boland PJ, et al. Estimating survival in patients with operable skeletal metastases: an application of a bayesian belief network. *PLoS One*. 2011;6(5):e19956.

7. Thio QCBS, Karhade AV, Bindels BJJ, et al. Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease. *Clin Orthop Relat Res*. 2020;478(2):322-33.

8. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63.

9. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *Int J Surg*. 2014;12(12):1495-9.

10. Wainer J. Comparison of 14 different families of classification algorithms on 115 binary datasets. *ArXiv*. 2016;abs/1606.00930.

11. Katagiri H, Okada R, Takagi T, et al. New prognostic factors and scoring system for patients with skeletal metastasis. *Cancer Med*. 2014;3(5):1359-67.

12. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.

13. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32-5.

14. Goutte C, Gaussier E. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *Springer*. 2005

15. Steyerberg EW, Van Calster B, Pencina MJ. Performance measures for prediction models and markers: evaluation of predictions and classifications. *Rev Esp Cardiol*. 2011;64(9):788-94.

16. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925-31.

17. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.

18. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. 1950;78(1):1-3.

19. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-8.

20. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14:40.

21. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk models was defined: from utopia

to empirical data. J *Clin Epidemiol.* 2016;74:167-76.

22. Bauer HC, Wedin R. Survival after surgery for spinal and extremity metastases. Prognostication in 241 patients. *Acta Orthop Scand.* 1995;66(2):143-6.

23. Janssen SJ, van der Heijden AS, van Dijke M, et al. 2015 Marshall Urist Young Investigator Award: prognostication in patients with long bone metastases: does a boosting algorithm improve survival estimates? *Clin Orthop Relat Res.* 2015;473(10):3112-21.

24. Nathan SS, Healey JH, Mellano D, et al. Survival in patients operated on for pathologic fracture: implications for end-of-life orthopedic care. *J Clin Oncol.* 2005;23(25):6072-82.

25. Willeumier JJ, van der Linden YM, van der Wal C, et al. An easy-to-use prognostic model for survival estimation for patients with symptomatic long bone metastases. *J Bone Joint Surg Am.* 2018;100(3):196-204.

26. Forsberg JA, Wedin R, Bauer HC, et al. External validation of the Bayesian Estimated Tools for Survival (BETS) models in patients with surgically treated skeletal metastases. *BMC Cancer.* 2012;12:493.

27. Ward E, Jemal A, Cokkinides V, et al. Cancer disparities by race/ethnicity and socioeconomic status. *CA Cancer J Clin.* 2004;54(2):78-93.

28. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol.* 2013;13(1):33.

29. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Stat Med.* 2000;19(24):3401-15.

CHAPTER

14

# INTERNATIONAL VALIDATION OF THE SORG MACHINE-LEARNING ALGORITHM FOR PREDICTING THE SURVIVAL OF PATIENTS WITH EXTREMITY METASTASES UNDERGOING SURGICAL TREATMENT

Ting-En Tseng*. Chia-Che Lee*, Hung-Kuan Yen*, Olivier Q. Groot, Chun-Han Hou, PShin-Ying Lin, Michiel ER. Bongers, Ming-Hsiao Hu, Aditya V. Karhade, Jia-Chi Ko, Yi-Hsiang Lai, Jing-Jen Yang, Jorrit-Jan Verlaan, Rong-Sen Yang, Joseph H. Schwab, Wei-Hsin Lin

**Joint first authorship*

*Clinical Orthopaedics and Related Research 2021*

# ABSTRACT

### Background

The Skeletal Oncology Research Group machine-learning algorithms (SORG-MLAs) estimate 90-day and 1-year survival in patients with long-bone metastases undergoing surgical treatment, and demonstrated good discriminatory ability in internal validation. However, the performance of a prediction model could potentially vary by race or region, and the SORG-MLA remains to be externally validated in an Asian cohort. Furthermore, the authors of the original developmental study did not consider the Eastern Cooperative Oncology Group (ECOG) performance status, a survival prognosticator repeatedly validated in other studies, in their algorithms because of missing data.

### Objectives

(1) Is the SORG-MLA generalizable to Taiwanese patients for predicting 90-day and 1-year mortality?

(2) Is the ECOG score an independent factor associated with 90-day and 1-year mortality while controlling for SORG-MLA predictions?

### Methods

All 356 patients who underwent surgery for long-bone metastases between 2014 and 2019 at one tertiary center in Taiwan were included. More than 98% (349/356) of the patients were of Han Chinese descent. A multivariate logistic regression analysis was used to evaluate whether the ECOG score was an independent prognosticator while controlling for the SORG-MLA's predictions – no retraining/recalibration was performed.

### Results

The SORG-MLAs had good discriminatory ability at both timepoints, with a c-index of 0.80 (95% confidence interval [CI], 0.74-0.86) for 90-day survival prediction and a c-index of 0.84 (95% CI, 0.80-0.89) for 1-year survival prediction. However, the calibration analysis showed that the SORG-MLAs tended to underestimate Taiwanese patients' survival (90-day survival prediction: calibration intercept, 0.78; 95% CI, 0.46-1.10; calibration slope, 0.74; 95% CI, 0.53-0.96; 1-year survival prediction: calibration intercept, 0.75; 95% CI, 0.49-1.00; calibration slope, 1.22; 95% CI, 0.95-1.49). The Brier score of the 90-day and 1-year SORG-MLA prediction models was lower than that of their respective null model (0.12 vs 0.16 for 90-day prediction; 0.16 vs 0.25 for 1-year prediction), indicating good overall performance of SORG-MLAs at these two timepoints. Decision curve analysis showed SORG-MLAs provided net benefits when threshold probabilities ranged from 0.40 to 0.95 for 90-day survival prediction and from 0.15 to 1.0 for 1-year prediction. The ECOG score was an independent factor associated with 90-day mortality (OR 1.94; 95% CI, 1.01-3.73) but not 1-year mortality (OR 1.07; 95% CI, 0.53-2.17) after controlling for SORG-MLA predictions for 90-day and 1-year survival, respectively.

**Conclusion**

Despite the baseline differences, the SORG ML algorithms retained good discriminative ability (area under the curve [AUC] 0.83; 95% confidence interval [CI] 0.76-0.88 for 90-day mortality and AUC 0.84; 95% CI 0.79-0.88 for 1-year mortality), calibration, overall performance, and decision curve analysis.

# INTRODUCTION

The incidence of long-bone metastases has been rising because of increased survival rates among patients with cancer.[1,2] Without proper treatment, a long-bone metastasis may cause skeleton-related events such as pain, disability, and pathologic fracture. These adverse events often lead to worse quality of life and are associated with higher mortality rates.[3,4] Commonly used nonoperative treatments for bone metastases include systemic chemotherapeutics, various types of radiation therapy, and bone-targeting agents such as bisphosphonates or denosumab. However, these treatment modalities rarely cure metastatic bone disease because of the aggressive nature of advanced-stage cancer, and surgical procedures may be indicated to address an impending or actual fracture of the involved bone.[2,5–8] It is challenging for clinicians to decide whether to offer surgical interventions for patients whose lifespans may be limited. Aside from the location of the metastasis, the extent of tumor involvement, response to adjuvant therapies, and severity of symptoms, the surgeon must also weigh the benefits, risks, and potential complications associated with surgery against the patient's expected survival.[9,10] Generally, patients with a short life expectancy may be treated nonoperatively if other means exist to properly control the local symptoms and maintain quality of life; or treated surgically with less invasive palliative techniques if they are not expected to have enough time to recover from a more extensive surgical procedure. Patients with longer expected survival are often given the choice of surgical procedures if other adjuvant therapies are deemed unlikely to achieve symptomatic relief or prevent fracture. These longer-term survivors may also benefit from tumor resection and more durable limb reconstruction, which achieves optimal local tumor control and sustained functional improvement. Two clinically practical time thresholds, namely 90-day (intermediate-term) and 1-year (long-term) survival, have been proposed for treatment decisions in patients with long-bone metastases.[9–11] Although patients who have not sustained a pathologic fracture and are expected to live fewer than 90 days are less likely to benefit from surgery, patients with an estimated survival of more than 1 year are candidates for more extensive surgery and durable reconstruction, such as prosthetic replacement.[12–17] An accurate survival estimation can thus help clinicians and patients in the shared decision-making process.

Several preoperative scoring systems have been developed to estimate patients' postoperative survival.[1,9–11,18–22] However, some of the scoring systems, such as the revised Katagiri score, did not achieve acceptable discriminatory ability in external validation.[1,20,23] Recently, Thio et al.[24] took advantage of the novel machine-learning concept and developed the Skeletal Oncology Research Group machine learning algorithm (SORG-MLA) to evaluate the intermediate-term and long-term survival probability of patients with extremity metastases. Although it has shown good discriminatory ability in the internal validation cohort of the developmental study, the SORG-MLA has not been externally validated.[25] Several studies suggested that racial distinctions among regions could influence the discriminatory ability of preoperative scoring systems because of differences in racial compositions, dominant cancer types, healthcare systems, and socioeconomic environments.[26–30] Han Chinese people account for 18% of the global population but constitute less than 5% of the US population.[31] In addition, several studies found that Chinese patients with certain types of malignancies had a better prognosis than their western counterparts.[5,32–36] In a world where international travel, education, and migration have become the norm, physicians in many countries could be seeing an increasing racially diverse patient population in their practice. It is therefore important to understand if a clinical tool such as SORG-MLAs can be generalized to different racial groups or used in regions outside of the United States.

The authors of the original SORG-MLA development study reported a lack of functional status data as one of their research limitations, and suggested future studies should include these factors to improve algorithm performance. The Eastern Cooperative Oncology Group (ECOG) performance scale is widely used by oncologists in clinical practice due to its simplicity, but not considered in the original development study due to missing data. It has also been shown to be associated with survival in cancer patients, while several preoperative scoring systems consider it as a prognosticator.[1,10,11,22,37] It would be of interest to know if ECOG should be investigated as a potential factor to be added into SORG-MLA to enhance the model's performance.

Therefore, in this study, we asked: (1) Is the SORG-MLA generalizable to a Taiwanese cohort for predicting 90-day and 1-year survival? (2) Is the ECOG score an independent factor associated with 90-day and 1-year survival while controlling for SORG-MLAs' predictions?

# METHODS

### Study Design and Setting

This international external validation study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.[38,39] The study was approved by our institutional review board (201912022RIND).

The selection criteria used in the development study were applied, resulting in 356 patients who underwent surgical treatment for long-bone metastases between 2014 and 2019 at the National Taiwan University Hospital (Figure 1).[24] In general, the indications for surgery were patients with an American Society of Anesthesiologists score of IV or below or patients considered fit for surgery based on a multidisciplinary assessment by a medical oncologist, anesthesiologist, and orthopedic surgeon; and the occurrence of a complete pathologic fracture or an impending pathologic fracture deemed unlikely to resolve with nonoperative treatment alone. Surgery was often offered for actual pathologic fractures of the femur unless clear medical contraindications existed such as ASA > 3, because femoral fractures tend to profoundly impact the patient's quality of life. An impending fracture was diagnosed if the lesion in question had a Mirels score ≥ 9 and caused pain or weakness in

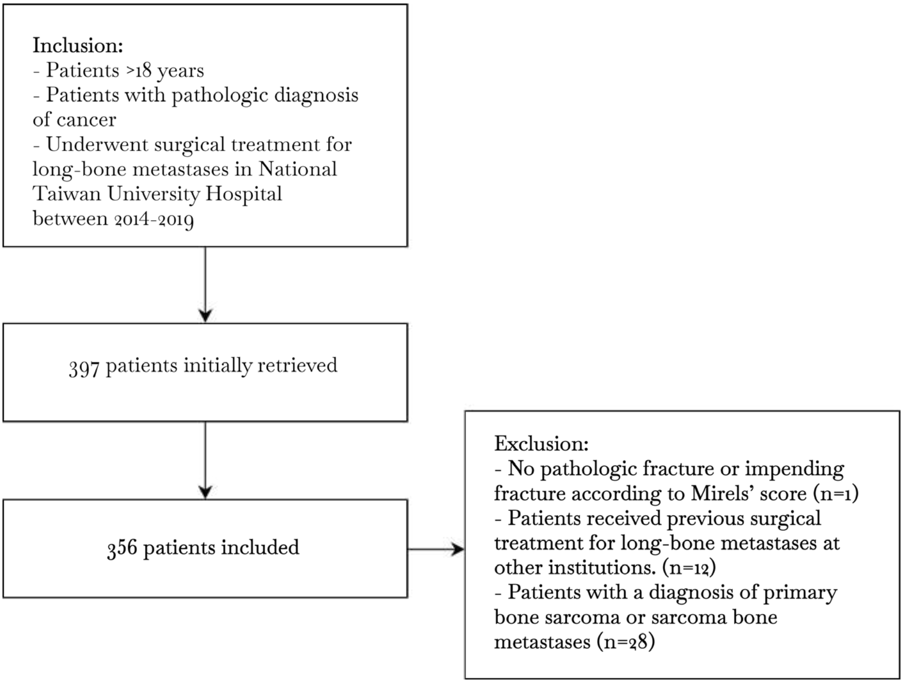

Inclusion:
- Patients >18 years
- Patients with pathologic diagnosis of cancer
- Underwent surgical treatment for long-bone metastases in National Taiwan University Hospital between 2014-2019
397 patients initially retrieved
356 patients included
Exclusion:
- No pathologic fracture or impending fracture according to Mirels' score (n=1)
- Patients received previous surgical treatment for long-bone metastases at other institutions. (n=12)
- Patients with a diagnosis of primary bone sarcoma or sarcoma bone metastases (n=28)

**Figure 1.** This flow diagram shows the enrolled patients.

the involved limb.[40] Patients with a diagnosis of primary bone sarcoma or sarcoma bone metastasis were excluded because these tumors include various histologic types and tend to behave differently than carcinomas do.[8,41,42]

### Participants' Baseline Characteristics

More than 98% (349/356) of the patients were of Han Chinese descent. The median age was 61 years (range 25-95 years), and 52% (184/356) of patients were women (Table 1). The median BMI was 26.6 kg/m$^2$ (range 13-39 kg/m$^2$). In this study, 27% (97/356) of patients had a slow-growth tumor, 33% (118/356) had a moderate-growth tumor, and 40% (141/356) had a rapid-growth tumor according to the definition proposed by Katagiri et al.[1] and later adopted in the original SORG-MLA development study.[24] In summary, hormone dependent breast cancer, hormone dependent prostate cancer, malignant lymphoma, malignant myeloma, and thyroid cancer were referred to as slow-growth tumors; non-small cell lung cancer with molecularly targeted therapy, hormone independent breast cancer, hormone independent prostate cancer, renal cell carcinoma, other gynecological cancer, and other cancers were referred to as moderate-growth tumors; other lung cancer, colon and rectal cancer, gastric cancer, hepatocellular carcinoma, pancreatic cancer, head and neck cancer, other urological cancer, esophageal cancer, malignant melanoma, gallbladder cancer, cervical cancer, and cancer of unknown origin were referred to as rapid-growth tumors. A pathologic fracture occurred in 55% (195/356) of patients, other-bone metastases were identified in 72% (256/356) of patients, visceral metastases were present in 51% (180/356) of patients, and brain metastases were found in 17% (60/356) of patients. Twenty-one percent (73/356) of patients had an ECOG score of 3 or 4. The most common surgical site was the lower extremities, in 76% (269/356) of patients. A total of 79% (281/356) of patients had pre-operative systemic medical therapy (defined as having at least one type of the following treatment: chemotherapy, targeted therapy, hormone therapy, or immunotherapy) and 60% (214/356) had local radiation. Six patients were lost to follow-up within 90 days, thirty were lost to follow-up within 1 year. Mortality at 90-days was 18% (63/350) and at 1-year 51% (167/326).

Baseline characteristics in the validation cohort differed from those in the original SORG-MLA development cohort reported by Thio et al.[24] in several regards (Table 1). The Taiwanese cohort had more patients with other Charlson comorbidities, moderate and rapid primary tumor growth, ECOG score of 3 or 4, preoperative systemic therapy, preoperative local radiation, and fewer other-bone metastases (all $p < 0.05$). The 90-day and 1-year mortality rates were higher in the developmental cohort than in the validation cohort (29% versus 18% and 62% versus 51%, respectively).

### Surgical Treatment

In general, stabilization with a nail or plate-and-screws construct followed by adjuvant radiotherapy was recommended for metastases from radio-sensitive tumors such as breast, prostate, lung cancer

**Table 1.** Baseline comparison between external validation (n=356) and development population (n=1090)

| *Variables* | *Validation (n=356)* | *Development (n=1090)* | |
|---|---|---|---|
| | *% (n) or median (range)* | | *P-value* |
| Age | 61 (25-95) | 63 (54-72) | 0.08 |
| Female sex | 52 (184) | 56 (610) | 0.16 |
| BMI (kg/m²) | 23 (13-39) | 27 (23-30) | |
| Other comorbidities | 60 (215) | 54 (584) | **0.03** |
| Histologic findings of the primary tumor | | | **<0.01** |
| Slow growth | 27 (97) | 42 (460) | |
| Moderate growth | 33 (118) | 24 (263) | |
| Rapid growth | 40 (141) | 34 (367) | |
| Primary tumor by location | | | |
| Lung | 33 (116) | 23 (247) | **<0.01** |
| Breast | 16 (58) | 24 (257) | **0.04** |
| Myeloma | 5 (18) | 15 (162) | **<0.01** |
| Renal | 6 (21) | 11 (117) | **<0.01** |
| Prostate | 5 (19) | 5 (58) | 0.99 |
| Lymphoma | 1 (5) | 4 (44) | **0.02** |
| Melanoma | 1 (2) | 3 (30) | **0.02** |
| Esophageal | 2 (7) | 2 (24) | 0.79 |
| Colon | 3 (10) | 2 (18) | 0.17 |
| Head and neck | 4 (16) | 2 (18) | **<0.01** |
| Thyroid | 2 (6) | 2 (18) | 0.97 |
| Other | 1 (5) | 2 (16) | 0.94 |
| Unknown | 2 (6) | 2 (16) | 0.77 |
| Pancreas | 2 (6) | 1 (7) | 0.07 |
| Sarcoma | 1 (1) | 1 (14) | 0.11 |
| Cervical | 1 (2) | 1 (1) | 0.09 |
| Other gynecologic | 1 (3) | 1 (13) | 0.58 |
| Other urologic | 3 (9) | 1 (12) | 0.05 |
| Hepatocellular carcinoma | 10 (36) | 1 (16) | **<0.01** |
| Stomach | 1 (4) | 1 (2) | **0.02** |
| Gallbladder | 2 (6) | 0 (0) | **<0.01** |
| Pathologic fracture | 55 (195) | 55 (594) | 0.93 |
| ECOG score | | | |
| 0-2 | 79 (283) | 85 (360) | **0.03** |
| 3-4 | 21 (73) | 15 (62) | |
| Tumor location | | | 0.69 |
| Upper extremity | 24 (87) | 23 (255) | |
| Lower extremity | 76 (269) | 77 (835) | |
| Other bone metastases | 72 (256) | 75 (845) | **0.03** |
| Visceral metastases | 51 (180) | 45 (487) | 0.05 |
| Brain metastases | 17 (60) | 16 (175) | 0.72 |
| Previous systemic therapy | 79 (281) | 62 (676) | **<0.01** |
| Local radiation | 60 (214) | 18 (194) | **<0.01** |

*Continued on next page*

| | | | |
|---|---|---|---|
| Preoperative laboratory values | | | |
| Hemoglobin level in g/dL | 11 (6-18) | 11 (10-13) | 0.18 |
| White blood cell count in $10^3$/uL | 7 (1-90) | 7 (5-10) | 0.93 |
| Platelet count in $10^3$/uL | 234 (36-651) | 251 (184-332) | 0.06 |
| Absolute lymphocyte count in $10^3$/uL | 1 (1-8) | 1 (1-2) | 0.48 |
| Absolute neutrophil count in $10^3$/uL | 5 (1-77) | 5 (4-8) | 0.86 |
| Neutrophil-to-lymphocyte ratio | 5 (1-67) | 5 (3-9) | 0.18 |
| Platelet-to-lymphocyte ratio | 216 (14-2776) | 234 (158-374) | 0.11 |
| Albumin level in g/dL | 4 (1-5) | 4 (3-4) | **<0.01** |
| ALP level in IU/L | 98 (23-2531) | 101 (74-146) | 0.10 |
| Calcium level in mg/dL | 9 (4-18) | 9 (9-10) | **<0.01** |
| Creatinine level in mg/dL | 0.7 (0.3-8.1) | 0.8 (0.7-1.1) | **<0.01** |
| Sodium level in mg/dL | 137 (118-149) | 138 (136-140) | **<0.01** |
| Mortality | | | |
| 90-day | 18 (63) | 29 (305) | **< 0.01** |
| 1-year | 51 (167) | 62 (639) | **< 0.01** |

*Baseline characteristics were compared using the chi-square test for categorical variables and Mann-Whitney U test for continuous variables.* Bold *indicates significance (p<0.05).*

[a] *BMI was missing for 0 patients in the validation cohort and for 22% (237 patients) of the developmental cohort. The ECOG score was missing for 0 patients in the validation cohort and 61% (668 patients) of the developmental cohort. Hemoglobin level was missing for 0 patients in the validation cohort and 13% (146 patients) of the developmental cohort. White blood cell count was missing for 0 patients in the validation cohort and 13% (146 patients) of the developmental cohort. Platelet count was missing for 0 patients in the validation cohort and 13% (146 patients) of the developmental cohort. The absolute lymphocyte count was missing for 2% (eight patients) of the validation cohort and 30% (326 patients) of the developmental cohort. The absolute neutrophil count was missing for 2% (eight patients) of the validation cohort and 30% (322 patients) of the developmental cohort. The albumin level was missing for 7% (25 patients) of the validation cohort and 30% (320 patients) of the developmental cohort. The alkaline phosphatase level was missing for 5% (18 patients) of the validation cohort and 20% (316 patients) of the developmental cohort. The calcium level was missing for 2% (eight patients) of the validation cohort and 18% (200 patients) of the developmental cohort. The creatinine level was missing for 0 patients in the validation cohort and 15% (66 patients) of the developmental cohort. The sodium level was missing for 1% (one patient) of the validation cohort and 18% (199 patients) of the developmental cohort.*

and hematologic malignancies. Metastatectomy and cement augmentation was typically performed for radio-resistant tumors such as renal cell carcinoma and hepatocellular carcinoma. Endoprosthetic replacement was considered for patients with an unsalvageable joint or extensive metaphyseal bone loss if they have a reasonably long survival, and for those who had oligometastatic disease and may benefit from wide excision of metastatic tumor. We tended to offer surgery to patients with actual femoral pathologic fractures even if their expected survival was shorter than 6 weeks, as non-surgical treatment in this setting rarely resulted in satisfactory pain control and improvement in quality of life. Fifty-nine percent (210/356) of patients were treated with intramedullary nailing, followed by plate-and-screws fixation in 23% (81/356), and endoprosthetic reconstruction in 18% (65/356).

### Outcomes and Explanatory Variables

The primary outcomes were 90-day and 1-year mortality, which were defined as the time between the patient's first surgery for a long-bone metastasis and death of any cause. Loss to follow-up occurred in 2% (6/356) patients at 90 days and in 8% (30/356) at 1 year. These patients whose final survival status could not be ascertained due to loss to follow-up were excluded from analyses of model performance and calculation of actual survival rates.

The following preoperative data were extracted: age; sex; BMI (in kg/m$^2$); any Charlson comorbidity in addition to metastatic cancer; primary tumor type, classified per Katagiri et al.[1]; ECOG score; tumor location; pathologic fracture; other bone, visceral (lung and/or liver), or brain metastases; previous systemic therapy or local radiation; absolute lymphocyte and neutrophil count (in 10$^3$/uL); albumin level (in g/dL), alkaline phosphatase level (in IU/L), calcium level (in mg/dL), creatinine level (in mg/dL), hemoglobin level (in g/dL), platelet count (in 10$^3$/uL), sodium level (in mg/dL), and white blood cell count (in 10$^3$/ uL).

### Statistical Analysis

We manually retrieved the 90-day and 1-year SORG-MLA predictions for each patient from an internet-based application (https://sorg-apps.shinyapps.io/extremitymetssurvival/). A discrimination analysis (concordance index [c-index]), calibration analysis (intercept and slope), overall performance analysis (Brier score) and decision curve analysis were performed to validate the two set of algorithms.[43,44] A C-index ranges from 0.5 to 1.0, with 0.5 indicating random guess and 1.0 perfect discrimination. A C-index ≥ 0.7 indicates a good model, and a C-index ≥ 0.8 an excellent model.[45] Calibration refers to the agreement between the predicted outcomes and the actual outcomes and is assessed by plotting the calibration curves and computing the calibration slope and intercept. A perfect calibration has an intercept of 0 and a slope of 1. A positive intercept suggests an underestimation of the outcome by the prediction model, and a negative intercept indicates an overestimation.[46] The Brier score refers to overall performance. It is the average mean squared difference between the model predictions and the observed outcomes, and ranges from 0 (best prediction) to 1 (worst prediction). However, the prevalence of the outcome must be considered; therefore, the Brier score of the null model was also calculated by assigning a probability equal to the prevalence of the outcome (in this case, the actual survival rate) to each patient. The net benefit of the prediction model is calculated by comparing its Brier score with that of the null model. If a prediction model's Brier score is lower than the null model's, then the prediction model is deemed as having good performance.

The decision curve analysis was designed to assess the net benefit of a model across a range of different threshold probabilities.[47] Unlike a discrimination analysis (c-index), a decision curve analysis considers the cost-to-benefit ratio. The user of the model can decide which threshold probability (i.e., the ratio of potential risk to the potential benefit) of a treatment is important or applicable, and determine if the model is valuable at that threshold and see what the predicted net benefit would be. In general, if the harm of a treatment modality is relatively limited – e.g., antibiotics for infection - the clinician may choose a lower threshold probability. In contrast, if the potential risks associated with a treatment are high, e.g., performing extensive surgery in a fragile patient, a higher threshold possibility should be chosen for decision-making.[44,48]

The baseline characteristics, 90-day mortality rate, and 1-year mortality rate of the developmental and external validation cohorts were compared. Continuous variables were assessed using one-way median tests. Categorical data were compared using chi-square tests and Yates's correction (if applicable). The actual and average predicted survival rates at 90 days and 1 year were compared with dependent t-test. A multivariate logistic regression analysis was fitted to the ECOG performance status to estimate 90-day and 1-year mortality while adjusting for the SORG-MLA prediction outcomes. The multivariate logistic regression results are provided as odds ratios (ORs) with 95% confidence intervals (CIs). The missForest method was used to impute missing values for the absolute lymphocyte count (2.2%; 8/356), absolute neutrophil count (2.2%; 8/356), albumin level (7.0%; 25/356), alkaline phosphatase level (5.0%; 18/356), calcium level (2.2%; 8/356), and sodium level (0.3%; 1/356).[49] No missing data was recorded for ECOG because the hospital's electronic medical records system requires input of ECOG score every time a patient with malignancy is seen in the clinic or admitted to the hospital. A two tailed p value ≤ 0.05 was considered significant. R for Mac (version 4.0.4; R Core Team), along with its packages of missForest, risk model decision analysis (RMDA), and CalibrationCurves (downloaded through Github), was used for all statistical analyses.

# RESULTS

## Performance

The SORG-MLAs showed good discriminatory ability in predicting the post-operative 90-day and 1-year survival in the Taiwanese cohort. The c-index was 0.80 (95% CI, 0.74-0.86; Table 2) for postoperative 90-day survival prediction and 0.84 (95% CI, 0.80-0.89; Table 2) for postoperative 1-year survival prediction. The calibration analysis provided an intercept of 0.78 (95%CI, 0.46-1.10) and slope of 0.74 (95% CI 0.53-0.96) for the 90-day survival prediction, and an intercept of 0.75 (95% CI, 0.49-1.00) and a slope of 1.22 (95% CI, 0.95-1.49) for 1-year survival (Figure 2). These positive calibration intercepts suggest that the SORG-MLAs tend to underestimate Taiwanese patients' survival at both post-operative 90 days and 1 year. The actual 90-day survival rate in our cohort was higher than the predicted value (82% vs 73%; dependent t test $p < 0.01$). The actual 1-year survival rate was also higher than the predicted 1-year survival rate (49% vs 35%; dependent t test ($p < 0.01$). The Brier score of the 90-day and 1-year SORG-MLA prediction models was lower than that of their respective null model (Table 2: 0.12 vs 0.16 for 90-day prediction; 0.16 vs 0.25 for 1-year prediction), indicating good overall performance of SORG-MLAs at these two timepoints. In the decision curve analysis, the 90-day SORG-MLA was shown to provide a positive net benefit compared with a strategy of operating on either all or no patients when the threshold probabilities ranged from 0.40 to 0.95 (Figure 3A). The 1-year SORG-MLA's also provided a similar gain of positive net benefit compared with a default

**Table 2.** C-indexes and Brier scores of the SORG-MLA by primary tumor histologic findings in the validation cohort (n=356)

| *Validation cohort* | *90-day prediction* | | | *1-year prediction* | | |
|---|---|---|---|---|---|---|
| | *C-index* | *Brier score* | *Actual vs predicted survival rate* | *C-index* | *Brier score* | *Actual vs predicted survival rate* |
| Overall (n=356) | 0.80 (0.74-0.86) | 0.12 (0.16) | 82% vs 73% | 0.84 (0.80-0.89) | 0.16 (0.25) | 49% vs 35% |
| Solid-organ (n=333) | 0.79 (0.73-0.86) | 0.13 (0.16) | 81% vs 72% | 0.84 (0.80-0.89) | 0.16 (0.25) | 47% vs 34% |
| Lung (n=116) | 0.87 (0.77-0.97) | 0.10 (0.16) | 82% vs 73% | 0.89 (0.83-0.95) | 0.13 (0.25) | 44% vs 34% |
| Breast (n=58) | 0.58 (0.16-1.00) | 0.07 (0.07) | 93% vs 83% | 0.75 (0.58-0.91) | 0.15 (0.17) | 78% vs 43% |
| Liver (n=37) | 0.72 (0.53-0.91) | 0.13 (0.14) | 85% vs 58% | 0.76 (0.58-0.94) | 0.18 (0.25) | 47% vs 26% |
| Hematologic (n=23) | 0.95[a] | 0.04 (0.05) | 96% vs 83% | 0.82 (0.58-1.00) | 0.15 (0.20) | 71% vs 49% |
| Kidney (n=21) | 0.65[a] | 0.05 (0.05) | 95% vs 80% | 0.80 (0.53-1.00) | 0.17 (0.24) | 39% vs 44% |
| Prostate (n=19) | 0.69 (0.43-0.94) | 0.21 (0.22) | 68% vs 77% | 0.98 (0.92-1.00) | 0.08 (0.24) | 42% vs 39% |

*The C-index is presented with 95% CIs in parentheses. A c-index of 0.5 indicates no better than random guess; that of 0.8 indicates a great discriminatory ability; that of 1 indicates the perfect discriminatory ability. The Brier score of the null model is presented in parentheses. The Brier score should be compared to the benchmark (Brier score of the null model which is presented in parentheses); a lower Brier score indicates a better overall performance of an algorithm. Solid-organ malignancies include all kind of malignancies except for hematopoietic malignancies.*

*[a]95% CI could not be calculated because only one patient died within 90 days of surgery.*

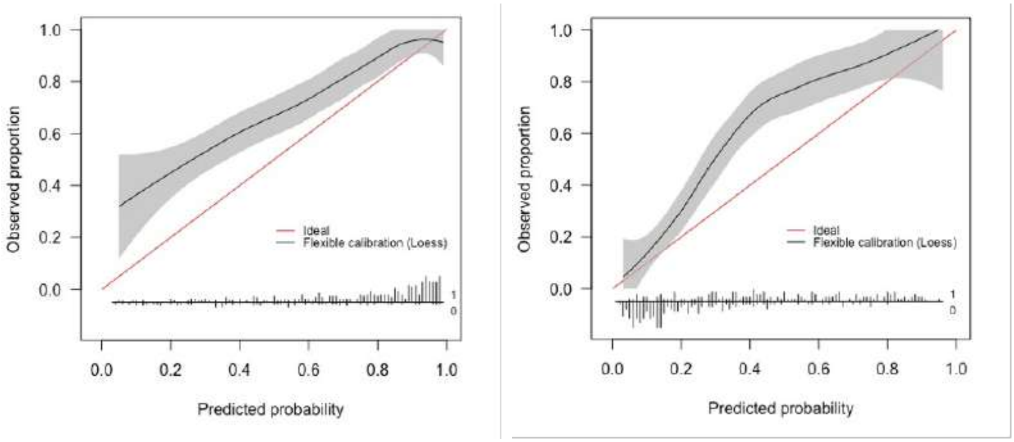


**Figure 2.** Calibration plots representing the predictions of the SORG-MLA are shown for (A) 90-day and (B) 1-year survival. The calibration plot visualizes how accurate the predictions are for different probabilities. The diagonal line represents the optimal calibration, the closer the line of the model, the more accurate the prediction. The calibration slope for the 90-day SORG-MLA was 0.74 (95% CI 0.46-1.10) and 1.22 (95% CI, 0.95-1.49) for the 1-year SORG-MLA. The calibration intercept for the 90-day SORG-MLA was 0.78 (95% CI, 0.46-1.10) and was 0.75 (95% CI, 0.49-1.00) for the 1-year SORG-MLA.

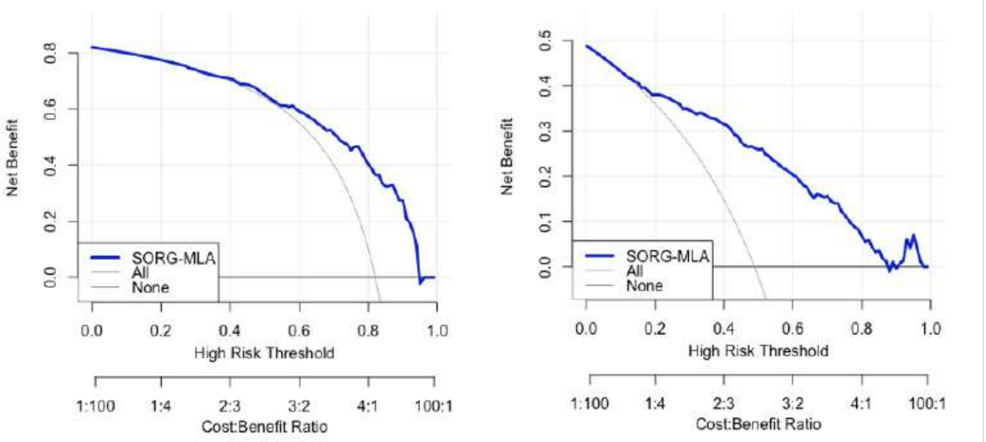


**Figure 3.** Decision curve analysis plots of predictions by the SORG-MLA are shown for (A) 90-day and (B) 1-year survival.

strategy of operating on either all or no patients when the threshold probabilities ranged from 0.15 to 1.0 (Figure 3B). These results indicated that management changes based on the 90-day and 1-year SORG algorithms had greater net benefit than the default strategies of changing management for no patients or for all patients.

### ECOG Score

The ECOG score was an independent factor associated with 90-day survival but not 1-year survival while controlling for SORG-MLAs' predictions. In the multivariate analysis that adjusted for SORG-MLA's 90-day survival prediction, patients with an ECOG score of 3 or 4 had higher 90-day mortality (OR 1.94; 95% CI, 1.01-3.73; p=0.04) but not 1-year mortality (OR 1.07; 95% CI, 0.53-2.17; p=0.85) than those with a score of 0-2.

## DISCUSSION

Patients with metastatic bone disease in the extremities should ideally be managed with a personalized strategy that takes their life expectancy into consideration to avoid under or over-treatment. The SORG-MLAs incorporate state-of-the-art machine learning techniques and have demonstrated excellent performance on internal validation. However, SORG-MLAs have not been externally validated outside the United States, especially in the Han Chinese population, who represent nearly one fifth of the global population. In this study, we found that SORG-MLAs retained excellent discriminatory ability and provided net benefits to surgical decision making when used to estimate both 90-day and 1-year survival probabilities in Taiwanese patients with extremity metastasis. However, the calibration analysis and a comparison of the actual and the predicted survival rates indicated that SORG-MLA tended to underestimate patient survival in our Taiwanese validation cohort. Clinicians should keep this under-estimation in mind when they use SORG-MLAs for survival prediction in patients of Han Chinese descent. The SORG-MLAs can be accessed online at https://sorg-apps.shinyapps.io/extremitymetssurvival/.

This study has several limitations. First, this was a single-institution study and more than 98% of the patients in our cohort were of Han Chinese descent. This might limit the reference value of the current study when physicians treat patients from other racially distinct regions. In addition, this cohort is unique because the Taiwanese healthcare system consists of the government-run National Health Insurance program, which covers every citizen and legal foreign resident, rendering molecularly targeted treatment and radiotherapy readily accessible and relatively affordable. For example, the price of gefitinib (Iressa), an effective targeted agent for lung cancer, is 10 times more expensive in the United States ($270 USD per tablet) than it is in Taiwan ($26 USD per tablet).[50,51] As a result, patients in Taiwan might be less financially constrained with use of newer medical therapies such as targeted agents and immunotherapy. Second, although we accounted for most known prognostic variables, additional factors – in particular tumor specific variables such as response to systemic therapy, use of oral targeted therapies or bone-modifying agents, administration of immunotherapy, and tumor molecular profiling– may be predictive of survival but were not included. Lack of consideration of these granular details could have contributed to the underestimation of patient survival in our validation cohort. We believe current predictive models can be improved upon not only by considering incremental factors such as the ECOG score identified in this study, but also by investigating the added value of these aforementioned variables. Third, this study is retrospective. The data used for input into the SORG-MLAs such as results of laboratory tests and variables based on imaging studies or clinical evaluation, were not acquired in a standardized fashion and not all at the same time before surgery. Validation of the SORG-MLAs based on data from prospectively enrolled cohort evaluated with a standardized pre-operative protocol is an avenue for future research. Fourth, survival is only one aspect to consider when deciding on surgical treatment. For example, some patients with femoral pathologic fracture might benefit from surgical fixation even though their expected survival is short because in this situation acceptable pain control and quality of life is seldom achieved with non-surgical treatment. Future studies should attempt to develop predictive models for outcomes such as postoperative ambulatory status, hospitalization, reoperations, systemic complications, level of pain, and quality of life, the latter of which is often considered to be the most important aspect in the care of patient with incurable cancer. Physicians should be aware of these potential pitfalls when using SORG-MLAs in the clinical setting.

In this study, we found that SORG-MLAs performed well in a cohort comprised mostly of Han Chinese, who represent a significant portion of the world's population and may be more and more frequently seen in many clinicians' practice in this age of globalization. This tool can help physicians and their Han Chinese patients in the shared decision-making process, but users should be aware that SORG-MLAs might under-estimate survival rates in this patient population. In a study comparing six state-of-the-art preoperative scoring systems for patients undergoing surgical treatment for long-bone metastases, Meares et al. reported that the PathFx model had the best performance for 90-day survival prediction (c-index 0.70; 95% CI, 0.69-0.70) and the OPTIModel was the best for predicting 1-year survival (c-index 0.79; 95% CI, 0.78-0.79).[23] Compared with these two benchmarks (the PathFx model and OPTIModel), the SORG-MLAs had better discriminatory ability at both timepoints (c-index 0.80; 95% CI, 0.74-0.86 for 90-day survival prediction and c-index 0.84; 95% CI 0.80-0.89 for 1-year survival prediction). However, PathFx was recently updated and has now been externally validated not only in patients treated with surgery but also in patients treated non-operatively with external beam radiation therapy.[9] In addition, PathFx provides postoperative survival predictions at six timepoints: 1 month, 3 months, 6 months, 12 months, 18 months, and 24 months. By contrast, the SORG-MLAs currently offer only 90-day and 1-year survival predictions and remain yet to be validated in non-operatively treated patients. The SORG-MLAs should ideally be retrained to make up for these shortcomings. Furthermore, cancer therapeutics have evolved and seen rapid advances in recent years. More emphasis is now placed on tumor-specific characteristics such as the histologic subtype, mutation status, hormone receptor expression profile, and response to novel treatment strategies. We believe future studies should focus on collecting granular tumor-specific data of individual cancer types to enhance the SORG-MLA's performance.

In our Taiwanese cohort, the ECOG performance scale was an independent factor associated with 90-day mortality but not with 1-year mortality after controlling for SORG-MLA predictions in multivariate analysis. This finding was consistent with results from several previous studies, in which investigators found that 90-day survival depended more on the patient's general condition (for example, the ECOG performance status or albumin level) and 1-year survival was influenced more by the primary tumor type.[24,28,29,52–54] One study specifically assigned quantified importance to various survival prognosticators for patients with spinal metastases.[28,29] On a scale of 0 to 100, where 100 indicated the most important prognosticators and 0 indicated the least important ones, the primary tumor type scored 100, the albumin level scored 90, and ECOG performance status scored less than 20 in one-year survival prediction. On the other hand, these three factors scored 60, 100, and 40, respectively, in 90-day survival prediction. We propose that developers of survival prediction algorithms should consider incorporating the ECOG score into their (machine learning) algorithms for predicting survival in patients with long-bone metastases. We believe that current predictive models can be improved upon by considering incremental factors such as the ECOG. Future studies should investigate the benefit of additional predictive factors such as tumor mutation profiles, novel systemic therapies, or body composition measurements based on imaging.[55]

# CONCLUSION

SORG-MLAs performed well in this Taiwanese cohort in terms of both discrimination and decision curve analysis. However, they tended to underestimate the patient's actual survival. The ECOG performance status may provide additional prognostic value for survival predictions, with further research warranted regarding this possibility. More international, larger-sized, and preferably prospective studies in search of additional prognosticators that add incremental value to the current model are needed to confirm and refine the findings of this study. The SORG-MLAs for extremity metastases can be accessed freely as an internet application at https://sorg-apps.shinyapps.io/extremitymetssurvival/.

# REFERENCES

1. Katagiri H, Okada R, Takagi T, et al. **New prognostic factors and scoring system for patients with skeletal metastasis.** *Cancer Med.* 2014;3(5):1359–1367.

2. Thio QCBS, Karhade AV, Ogink PT, et al. **Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease.** *Clin Orthop Relat Res.* 2020;478(2):322–333.

3. Coleman RE. **Metastatic bone disease: clinical features, pathophysiology and treatment strategies.** *Cancer Treat Rev.* 2001;27(3):165–176.

4. Roodman GD. **Mechanisms of bone metastasis.** *Discov Med.* 2004;4(22):144–148.

5. Shieh SH, Hsieh VCR, Liu S-H, et al. **Delayed time from first medical visit to diagnosis for breast cancer patients in Taiwan.** *J Formos Med Assoc.* 2014;113(10):696–703.

6. Tattersall MH, Thomas H. **Recent advances: oncology.** *BMJ.* 1999;318(7181):445–448.

7. Wells A, Grahovac J, Wheeler S, et al. **Targeting tumor cell motility as a strategy against invasion and metastasis.** *Trends Pharmacol Sci.* 2013;34(5):283–289.

8. Yap YS, Lu YS, Tamura K, et al. **Insights into breast cancer in the east vs the west: a review.** *JAMA Oncol.* 2019;5(10):1489–1496.

9. Anderson AB, Wedin R, Fabbri N, et al. **External validation of PATHFx version 3.0 in patients treated surgically and nonsurgically for symptomatic skeletal metastases.** *Clin Orthop Relat Res.* 2020;478(4):808–818.

10. Forsberg JA, Eberhardt J, Boland PJ, et al. **Estimating survival in patients with operable skeletal metastases: an application of a bayesian belief network.** *PLoS One.* 2011;6(5):e19956.

11. Sørensen MS, Gerds TA, Hindsø K, et al. **Prediction of survival after surgery due to skeletal metastases in the extremities.** *Bone Joint J.* 2016;98-B(2):271–7.

12. Ruggieri P, Mavrogenis AF, Casadei R, et al. **Protocol of surgical treatment of long bone pathological fractures.** *Injury.* 2010;41(11):1161–1167.

13. Bickels J, Dadia S, Lidar Z. **Surgical management of metastatic bone disease.** *J Bone Joint Surg Am.* 2009;91(6):1503–1516.

14. Harvey N, Ahlmann ER, Allison DC, et al. **Endoprostheses last longer than intramedullary devices in proximal femur metastases.** *Clin Orthop Relat Res.* 2012;470(3):684–691.

15. Scotti C, Camnasio F, Peretti GM, et al. **Modular prostheses in the treatment of proximal humerus metastases: review of 40 cases.** *J Orthop Traumatol.* 2008;9(1):5–10.

16. Steensma M, Boland PJ, Morris CD, et al. **Endoprosthetic treatment is more durable for pathologic proximal femur fractures.** *Clin Orthop Relat Res.* 2012;470(3):920–926.

17. Wedin R. **Surgical treatment for pathologic fracture.** *Acta Orthop Scand Suppl.* 2001;72(302):2p., 1–29.

18. Bauer HC, Wedin R. **Survival after surgery for spinal and extremity metastases. Prognostication in 241 patients.** *Acta Orthop Scand.* 1995;66(2):143–146.

19. Janssen SJ, van der Heijden AS, van Dijke M, et al. **2015 Marshall Urist Young Investigator Award: prognostication in patients with long bone metastases: does a boosting algorithm improve survival estimates?** *Clin Orthop Relat Res.* 2015;473(10):3112–21.

20. Nathan SS, Healey JH, Mellano D, et al. Survival in patients operated on for pathologic fracture: implications for end-of-life orthopedic care. *J Clin Oncol.* 2005;23(25):6072–6082.

21. Ratasvuori M, Wedin R, Keller J, et al. Insight opinion to surgically treated metastatic bone disease: Scandinavian Sarcoma Group Skeletal Metastasis Registry report of 1195 operated skeletal metastasis. *Surg oncol.* 2013;22(2):132–8.

22. Willeumier JJ, van der Linden YM, van der Wal CWPG, et al. An easy-to-use prognostic model for survival estimation for patients with symptomatic long bone metastases. *J Bone Joint Surg Am.* 2018;100(3):196–204.

23. Meares C, Badran A, Dewar D. Prediction of survival after surgical management of femoral metastatic bone disease - A comparison of prognostic models. *J Bone Oncol.* 2019;15:100225.

24. Thio QCBS, Karhade AV, Bindels BJJ, et al. Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease. *Clin Orthop Relat Res.* 2020;478(2):322–333.

25. Groot OQ, Bindels BJJ, Ogink PT, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop.* 2021:1–9.

26. Bongers MER, Karhade AV, Villavieja J, et al. Does the SORG algorithm generalize to a contemporary cohort of patients with spinal metastases on external validation? *Spine J.* 2020;20:1646-1652.

27. Karhade AV, Ahmed AK, Pennington Z, et al. External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease. *Spine J.* 2020;20(1):14–21.

28. Karhade AV, Thio QCBS, Ogink PT, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. *Neurosurgery*. 2019;1:85;671-681

29. Shah AA, Karhade AV, Park HY, et al. Updated external validation of the SORG machine learning algorithms for prediction of ninety-day and one-year mortality after surgery for spinal metastasis. *Spine J.* 2021:31;1529-9430

30. Yang JJ, Chen CW, Fourman MS, et al. International external validation of the SORG machine learning algorithms for predicting 90-day and 1-year survival of patients with spine metastases using a Taiwanese cohort. *Spine J.* 2021:2;1529-9430

31. Zhao Y-B, Zhang Y, Zhang QC, et al. Ancient DNA reveals that the genetic structure of the northern Han Chinese was shaped prior to 3,000 years ago. *PLoS One*. 2015;10(5):e0125676.

32. Chang C-W, Tai H-C, Cheng N-C, et al. Risk factors for complications following immediate tissue expander based breast reconstruction in Taiwanese population. *J Formos Med Assoc.* 2017;116(1):57–63.

33. Chen C-H, Lu Y-S, Cheng A-L, et al. Disparity in tumor immune microenvironment of breast cancer and prognostic impact: asian versus western populations. *Oncologist*. 2020;25(1):e16–e23.

34. Chen C-H, Tzai T-S, Huang S-P, et al. Clinical outcome of Taiwanese men with metastatic prostate cancer compared with other ethnic groups. *Urology*. 2008;72(6):1287–1292.

35. Park J-W, Chen M, Colombo M, et al. Global patterns of hepatocellular carcinoma management from diagnosis to death: the BRIDGE Study. *Liver Int*. 2015;35(9):2155–2166.

36. Wu C-E, Chen S-C, Chang H-K, et al. Identification of patients with hormone receptor-positive breast cancer who need adjuvant tamoxifen therapy for more than 5 years. *J Formos Med Assoc.* 2016;115(4):249–256.

37. Kantarjian H, O'brien S, Cortes J, et al. Results of intensive chemotherapy in 998 patients age 65 years or older with acute myeloid leukemia or high-risk myelodysplastic syndrome: predictive prognostic models for outcome. *Cancer*. 2006;106(5):1090–1098.

38. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for

individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* 2015;13:1.

39. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol.* 2008;61(4):344–349.

40. Mirels H. Metastatic disease in long bones. A proposed scoring system for diagnosing impending pathologic fractures. *Clin Orthop Relat Res.* 1989;(249):256–264.

41. Nystrom LM, Reimer NB, Reith JD, et al. Multidisciplinary management of soft tissue sarcoma. *Scientific World Journal.* 2013;2013:852462.

42. San-Julian M, Diaz-de-Rada P, Noain E, et al. Bone metastases from osteosarcoma. *Int Orthop.* 2003;27(2):117–120.

43. Fredon A, Radchenko AK, Cuppen HM. Quantification of the role of chemical desorption in molecular clouds. *Acc Chem Res.* 2021;54(4):745–753.

44. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35(29):1925–1931.

45. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol.* 2010;5(9):1315–1316.

46. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17(1):230.

47. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak.* 2006;26(6):565–574.

48. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21(1):128–138.

49. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics.* 2012;28(1):112–118.

50. Aguiar PNJ, Haaland B, Park W, et al. Cost-effectiveness of osimertinib in the first-line treatment of patients with GFR-mutated advanced non-small cell lung cancer. *JAMA Oncol.* 2018;4(8):1080–1084.

51. Hsu JC, Wei C-F, Yang S-C. Effects of removing reimbursement restrictions on targeted therapy accessibility for non-small cell lung cancer treatment in Taiwan: an interrupted time series study. *BMJ Open.* 2019;9(3):e022293.

52. Karhade AV, Thio QCBS, Ogink PT, et al. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. *Neurosurgery.* 2019;85(1):E83–E91.

53. Massaad E, Shin JH. Commentary: sarcopenia as a prognostic factor for 90-day and overall mortality in patients undergoing spine surgery for metastatic tumors: a multi-center retrospective cohort study. *Neurosurgery.* 2020;87(5):E550–E551.

54. Pielkenrood BJ, van Urk PR, van der Velden JM, et al. Impact of body fat distribution and sarcopenia on the overall survival in patients with spinal metastases receiving radiotherapy treatment: a prospective cohort study. *Acta Oncol.* 2020;59(3):291–297.

55. Kapoor ND, Twining PK, Groot OQ, et al. Adipose tissue density on CT as a prognostic factor in patients with cancer: a systematic review. *Acta Oncol.* 2020;59(12):1488–1495.

CHAPTER

15

# COMPARISON OF SKELETAL ONCOLOGY RESEARCH GROUP (SORG) CLASSICAL ALGORITHM AND MACHINE LEARNING ALGORITHMS FOR SURVIVAL ESTIMATION IN SPINAL METASTASIS: A META-ANALYSIS OF THE LITERATURE

Hung-Kuan Yen*, Wei-Hsin Lin*, Olivier Q. Groot, Chih-Wei Chen, Michiel E.R. Bongers, Aditya V. Karhade, Akash A. Shah, Jiun-Jen Yang, Tse-Chuan Yang, Bas J.J. Bindels, Jorrit-Jan Verlaan, Joseph H. Schwab, Shu-Hua Yang, Francis J. Hornicek, Ming-Hsiao Hu

**Joint first authorship*

*The Spine Journal (Under review)*

# ABSTRACT

### Background

The Skeletal Oncology Research Group (SORG) classical algorithm (CA) and machine learning algorithms (MLA) provide good discrimination for postoperative 90-day and 1-year survival of spinal metastatic disease. However, one non-American validation of SORG-MLA demonstrated diminished performance results, indicating racially or geographically distinct regions should be considered to ensure accurate and reliable prognoses.

### Objective

Using a meta-analysis to determine the pooled discriminatory ability of both algorithms; and (2) test the hypothesis SORG-CA has less variability in performance than SORG-MLA in non-American validation cohorts as SORG-CA does not incorporates regional-specific variables such as BMI as input.

### Design

Systematic review and meta-analysis.

### Methods

This meta-analysis included seven studies with five similar cohorts each for SORG-CA and SORG-MLA, of which only one cohort was non-American (Taiwanese). Pooled logit-transformed area under receiver operating characteristic curves (logit (AUC)) overall and by region (America vs. non-America) were provided for both algorithms at both time points.

### Results

The pooled logits (AUC)s were 0.82, 1.11, 1.36, and 1.57 (95% CI, 0.53-0.11, 0.74-1.48, 1.09-1.63, 1.17-1.98, respectively) for 90-day and 1-year SORG-CA, 90-day and 1-year SORG-MLA, respectively. All the algorithms performed better in United States than in Taiwan (P<0.001). The performance of SORG-CA was less influenced by a non-American cohort than SORG-MLA.

### Conclusion

These observations might highlight the importance of incorporating region-specific variables into existing models to make them generalizable to racially or geographically distinct regions.

# INTRODUCTION

Approximately two-thirds of cancer patients eventually develop bony metastasis, with the spine being the most common site.[40] Surgery is often needed in patients with symptomatic spinal metastasis to improve survival, relieve pain, restore spinal stability, prevent neurologic decline, and improve or at least maintain quality of life.[33,36] The benefits of surgery should be balanced with the possible drawbacks such as complications or premature death.[13,31] Many algorithms have been developed to predict patients' survival; however, their accuracy has not been consistent, or they have not been validated with racially diverse data as compared with the developmental cohort which is predominantly white population.[24,37,47,54] A meta-analysis could save time from multiple repeat validations and provide a more comprehensive picture of the status quo, potentially leading to improved survival predictions in racially distinct patients from the developmental cohort.

The Skeletal Oncology Research Group (SORG) machine learning algorithms (MLA) proposed by Karhade et al. demonstrated a high accuracy of mortality rate estimation at 90 days and 1 year.[21] Four external validation studies lent support to this algorithm.[8,20,39,50] However, three of the cohorts were highly homogenous in their population makeup and had similar healthcare systems since they were all in the eastern United States,[8,20,21] where individuals having origins from Asia, only account on average for 5% of the population.[3] Yang et al. conducted the first external validation in a Taiwanese population of 427 patients to verify if the SORG-MLA maintained its discriminatory ability in a different racial group.[50] The algorithms exhibited good discrimination but tended to systematically underestimate survival in Taiwanese patients.[50] Yang et al. hypothesized that the underestimation might be resulted from the difference in the baseline body mass index (BMI) between the Taiwanese and American populations (Table 1) and suggested that region-specific BMI should be considered in predicting survival of patients with spinal metastasis.[50]

Before adopting the SORG-MLA, Pereira et al. proposed the SORG classical algorithm (SORG-CA) to estimate patients' 90-day and 1-year survival.[34] It also retained good discriminatory power in three validation cohorts from US.[5,35,39] While most of the parameters included in the SORG-CA and SORG-MLA were similar or not region-specific, SORG-CA did not include BMI as a prognosticator (Table 1). If region-specific BMI contributes to the less accurate prediction of patient survival by SORG-MLA in the Asian population, as proposed by Yang et al., the predictive power of SORG-CA might have less variability. Therefore, we sought to externally validate the SORG-CA with the Taiwanese patient cohort of 366 patients and compared the performance of SORG-CA with SORG-MLA to test Yang's hypothesis.

The current study aimed to (1) perform a meta-analysis to determine the pooled accuracy of survival prediction by the SORG-CA and SORG-MLA; (2) test the hypothesis that the performance of SORG-CA outperforms SORG-MLA when applied to the non-American validation cohorts.

**Table 1.** SORG-CA and SORG-MLA input parameters

| | SORG-CA | SORG-MLA |
|---|---|---|
| **Predictors** | **Pereira, 2016** | **Karhade, 2019** |
| Patient background | Age≥65: yes vs. no<br>ECOG: 0-2 vs. 3-4 | Body mass index (kg/m$^2$)<br>ECOG: 0-2 vs. 3-4<br>ASIA: A-D vs. E<br>Additional Charlson comorbidity other than metastatic disease |
| Primary tumor site* | 1. Best prognosis: lymphoma, breast cancer, multiple myeloma, kidney cancer, prostate cancer, or thyroid cancer<br>2. Worse prognosis: others | 1. Best prognosis: hormone dependent breast cancer and prostate cancer, malignant lymphoma, malignant myeloma, thyroid cancer<br>2. Moderate prognosis: NSCLC with molecularly targeted therapy, hormone independent breast cancer and prostate cancer, RCC, sarcoma, other gynecological cancer, others<br>3. Worst prognosis: other lung cancer, colon and rectal cancer, gastric cancer, HCC, pancreatic cancer, head and neck cancer, other urological cancer, esophageal cancer, malignant melanoma, gallbladder cancer, cervical cancer, unknown origin |
| Tumor status and treatment | Previous systemic therapy: yes vs. no<br>Visceral (lung/liver) metastases: yes vs. no<br>Presence of multiple spine metastases: yes vs. no<br>Brain metastases: yes vs. no | Prior systemic therapy: yes vs. no<br>Visceral metastases: yes vs. no<br>Number of spine metastases: 1-2 vs. ≥3<br>Brain metastases: yes vs. no |
| Laboratory tests | White blood cell count≥11,000/mm$^3$: yes vs. no<br>Hemoglobin≤10 g/dL: yes vs. no | Absolute lymphocyte count (×10$^3$/ L),<br>Absolute neutrophil count (×10$^3$/ L),<br>Hemoglobin (g/dL)<br>Platelet count (×10$^3$/ L),<br>Alkaline phosphatase (IU/L),<br>Albumin (g/dL)<br>Creatinine (mg/dL)<br>International Normalized Ratio |

*SORG=Skeletal Oncology Research Group; CA=classical algorithm; MLA=machine learning algorithm; ECOG=Eastern Cooperative Oncology Group; ASIA=American Spinal Injury Association; NSCLC=Non-small cell lung cancer; RCC=Renal cell carcinoma; HCC=Hepatic cell carcinoma.*

** Difference in categorization exist in different definition (Pereira used the Katagiri categorization proposed in 2005, Karhade used the updated Katagiri categorization proposed in 2015)*

# METHODS

### Guidelines

This study follows the Meta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines, the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement for reporting meta-analysis, and the Declaration of Helsinki and Health Insurance Portability and Accountability Act guidelines.[27,44] This study was not registered ahead. No ethical approval was needed for conducting a systematic review of published literature. Electric-medical-records review for external validation was approved by our institutional review board (2020005016RINC).

### Search Strategy and Selection Criteria

This meta-analysis followed a detailed, prespecified protocol, which set out the objectives, inclusion criteria for cohort studies, data to be collected, and analyses to be done (available on request). Studies considered for the meta-analysis had to include only adult patients (>18 years) with spinal metastasis who underwent surgery and were followed up for at least 1-year. On the other hand, studies with more than 30% missing for each parameter were excluded.[42] The proportion of missing data and ways to handle them should be disclosed. Second, studies not reporting clinicodemographic comparison with the developmental cohort were excluded. Third, any study not presenting discrimination analysis, in terms of either area under the receiver operating characteristic curve (AUC) or odds ratio (OR), was also excluded.[48] The corresponding 95% confidence intervals (CI) should also be given.

Three databases (PubMed, Embase, and Google scholar) were searched as of March 30th 2021, and the searching terms were provided in Appendix Table 1. The corresponding MeSH terms were also searched. In addition, we manually retrieved all articles that cited the two articles proposing SORG-CA and SORG-MLA.[21,34] Unpublished studies were also initially included to avoid publication bias. Whereas, no unpublished studies were finally analyzed due to overlapping participants with other published studies or did not meet the inclusion criteria, and therefore they were not eventually analyzed. As several studies suggested that including non-English publications does not change the conclusions,[30,32] we opted to only include English studies. Two researchers independently screened all studies that were identified through the literature search and evaluated full-text studies using the predefined criteria. Disagreements were solved by a discussion with another author.

### Methodological Quality Assessment

To avoid potential bias, the methodological quality of included studies was assessed by two independent authors, of which the latter was not affiliated with any of the institutions from where the studies originated. The studies were assessed following Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines and the Prediction model Risk Of Bias ASsessment Tool (PROBAST).[12,49] The PROBAST tool by the Cochrane Prognosis group determines risk of bias and applicability of prediction models in systematic reviews. This tool consists of 20 signaling questions across four domains: participants selection (1), predictors (2), outcome (3), and analysis (4). Each domain is rated "low," "high," or "unclear" (Appendix Figure 1). All studies had low risk of bias and one study was rated "unclear" in the analysis section because they did not provide the standard error of AUC of the validation data (Table 2).[21] However, we included the study in our quantitative analysis after confirming with the authors' paper that the validation and training set shared great homogeneity since patients were randomly assigned.

### Data Extraction

Two authors independently extracted data from the studies included in quantitative synthesis using standardized sheets. The items were: first author, year published, enrolment period, country, sample size, level of evidence, survival proportion at 90-day and 1-year, and the discrimination of the algorithms in terms of either AUC or OR and its 95% CI. A third reviewer was consulted to resolve any disagreements in the extracted data. Although we have contacted all corresponding authors, we could only estimate the proportion of Asian population in the studies due to non respondence or their privacy policy. The included SORG authors were not part of the data extraction.

We retrieved all consecutive 427 adult patients for the external validation of SORG-CA, which were already used to externally validate the SORG-MLA, with spinal metastasis that underwent surgical treatment between November 1st 2010, and December 31st 2018 at our tertiary center.[50] Sixty-one (14%) patients were excluded due to missing data of clinical presentation, imaging, laboratory values, or operative documents as the SORG-CA model requires complete data (Appendix Figure 2). Finally, 366 patients were included for validating SORG-CA.

### Outcomes and Explanatory Variables

The primary outcomes were the pooling logit-transformed AUC (logit(AUC)) with 95% CI of SORG-CA and SORG-MLA. An AUC of 1suggests the best discriminatory ability and has a corresponding logit(AUC) approaching infinity; the worst AUC of 0 indicates the discriminatory ability is no better than random guessing, and has a corresponding logit(AUC) of 0. A logit(AUC) of 0.85, equivalent to an AUC of 0.7, was generally considered the cut-off value for designating a prediction model as having an acceptable discriminatory ability; a logit(AUC) of 2.20, equivalent to an AUC of 0.9, was viewed as an indicator of outstanding discriminatory ability of a model. Since logit(AUC), instead of AUC itself, had a normal distribution and could be pooled easily for meta-analysis, this transformation could circumvent the bounded nature of the AUC. The secondary outcome was the performance expressed in logit(AUC) with 95% CI of SORG-CA compared with SORG-MLA using the similar Taiwanese external validation cohort. The subgroup analysis was also conducted by using a meta-regression.

The following variables were included based on necessary input for SORG-CA: age (years), Eastern Cooperative Oncology Group [ECOG] performance status, primary tumor histology, presence of visceral metastasis (metastasis in liver or lung), presence of brain metastasis, presence of spine metastasis, previous systemic therapy, and preoperative white blood cell (WBC) count (x10$^3$per microliter) and hemoglobin level (grams per deciliter [g/dL]) (Table 1).[34] The outcome of SORG-CA was the discrimination based on odds ratio.

### Statistical Analysis

The extracted ORs and their 95% CI were converted into AUCs and corresponding CIs.[48] Then the logit-transformation was performed for further analysis.[7,14] For subgroup analysis, two logit(AUC) s were compared by Student's t-test since the logit(AUC) had a normal distribution.[11,14] We also conducted subgroup analysis by meta-regression. Multivariable regression was conducted using a backward stepwise procedure, and P-value was set at 0.10 for entry and at 0.20 for exit. Fixed effect size was referenced unless a high level of heterogeneity (i.e., $I^2$>50%) was observed. Funnel plots were not given due to the small numbers of pooling literatures. The clinicodemographic data were compared with developmental cohorts by either chi-square tests or Student's t-test. Statistically significant level was set at 0.05. We used R for Mac (version 4.0.2; R Core Team. St. Louis) for data analysis.

## RESULTS

### Search Results and Characteristics

After screening 121 studies, we assessed 25 full-text studies for eligibility, and ultimately seven studies were included in quantitative synthesis after critical appraisal (Figure 1). In total, five different cohorts were used to validate the studies, four of which originated from the US (Table 2 and Figure 2). Only one non-American cohort was included. The AUCs of SORG-CA ranged from 0.61 to 0.78 for 90-day and 0.65 to 0.85 for 1-year survival. The AUCs of SORG-MLA ranged from 0.73 to 0.84 for 90-day and 0.74 to 0.90 for 1-year survival. Of notice, the best AUCs of both algorithms were reported in an external validation cohort from the west coast of the US.

### Overall Discrimination of SORG-CA  SORG-MLA

In SORG-CA, the pooling logit(AUC)s showed great overall discrimination for 90-day (logit(AUC), 0.82; 95% CI, 0.53-1.11; Figure 3A) survival in 1,641 patients and 1-year (logit(AUC), 1.11; 95% CI, 0.74-

**Table 2.** Summary of included studies

| *Author* | *Characteristics of studies* | | | | |
|---|---|---|---|---|---|
| | *Year* | *Study period* | *Country* | *Institution* | *Method for missing data* |
| SORG -CA | | | | | |
| Pereira | 2017 | 2014 | USA | MSK | No missing data |
| Ahmed | 2018 | 2003-2016 | USA | JHH | No missing data |
| Karhade | 2019 | 2000-2016 | USA | MGH, BWH | MissForest methodology |
| Shah | 2021 | 2004-2020 | USA | UCLA | MissForest methodology |
| Yen | Current study | 2010-2018 | Taiwan | NTUH | No missing data |
| SORG-MLA | | | | | |
| Karhade | 2019 | 2000-2016 | USA | MGH, BWH | MissForest methodology |
| Bongers | 2020 | 2014-2016 | USA | MSK | MissForest methodology |
| Karhade | 2020 | 2003-2016 | USA | JHH | MissForest methodology |
| Yang | 2021 | 2010-2018 | Taiwan | NTUH | MissForest methodology |
| Shah | 2021 | 2004-2020 | USA | UCLA | MissForest methodology |

*AUC=area under receiver operating characteristic curves; CI=confidence interval; SORG=Skeletal Oncology Research Group; CA=classical algorithm; USA=United States of America; MSK=Memorial Sloan Kettering Cancer Center; JHH=Johns Hopkins Hospital; MGH=Massachusetts General Hospital; BWH=Brigham and Women's Hospital; UCLA=University of California, Los Angeles; NTUH=National Taiwan University Hospital; MLA=machine learning algorithm.*

1.48; Figure 3B) survival in 1,598 patients. The corresponding AUCs were 0.69 for 90-days and 0.75 for 1-year survival. In SORG-MLA, the pooling logit(AUC)s also showed great overall performances for 90-day (logit(AUC), 1.36; 95% CI, 1.09-1.63; Figure 3C) survival in 1,796 patients and 1-year (logit(AUC), 1.57; 95% CI, 1.17-1.98; Figure 3D) survival in 1,753 patients. The corresponding AUCs were 0.80 for 90 day and 0.83 for 1-year survival. The $I^2$s of the four pooling logit(AUC)s (including fixed effect and random effect) showed great heterogeneity (all P<0.01). SORG-MLA performed better than the SORG-CA (P<0.001; Appendix Table 2). Sensitivity analysis excluding developmental and internal validation cohort showed similar results that both algorithms provided great discriminatory ability and SORG-MLA had a better performance (Appendix Figure 5).

### Pooled Discrimination by Regions

In SORG-CA, the pooled results using the four US cohorts showed great discriminatory ability for 90-day SORG-CA (logit(AUC), 0.90; 95% CI, 0.56-1.24; Appendix Figure 4A) and 1-year (logit(AUC),

| *Characteristics of patients* | | | | |
|---|---|---|---|---|
| *Time period* | *Sample size* | *Mortality (%)* | *AUC* | *95% CI* |
| 90-day | 100 | 21 (21%) | 0.61 | 0.55-0.68 |
| 1-year | 176 | 56 (56%) | 0.65 | 0.57-0.74 |
| 90-day | 176 | 40 (27%) | 0.67 | 0.57-0.77 |
| 1-year | 176 | 99 (56%) | 0.77 | 0.70-0.84 |
| 90-day | 722 | 181 (25%) | 0.73 | 0.69-0.77 |
| 1-year | 709 | 385 (54%) | 0.76 | 0.72-0.80 |
| 90-day | 277 | 61 (22%) | 0.78 | 0.72-0.84 |
| 1-year | 253 | 133 (53%) | 0.85 | 0.81-0.89 |
| 90-day | 366 | 105 (29%) | 0.64 | 0.57-0.70 |
| 1-year | 360 | 347 (69%) | 0.69 | 0.61-0.79 |
| | | | | |
| 90-day | 722 | 181 (25%) | 0.83 | 0.81-0.85* |
| 1-year | 709 | 385 (54%) | 0.86 | 0.84-0.88* |
| 90-day | 199 | 55 (28%) | 0.81 | 0.74-0.87 |
| 1-year | 197 | 124 (63%) | 0.84 | 0.77-0.89 |
| 90-day | 176 | 40 (23%) | 0.75 | 0.66-0.83 |
| 1-year | 176 | 99 (56%) | 0.77 | 0.70-0.83 |
| 90-day | 422 | 105 (25%) | 0.73 | 0.67-0.78 |
| 1-year | 418 | 247 (59%) | 0.74 | 0.69-0.79 |
| 90-day | 277 | 61 (22%) | 0.84 | 0.79-0.89 |
| 1-year | 253 | 133 (53%) | 0.90 | 0.86-0.93 |

** This data was calculated assuming the research remained homogenous between training and validation sets for the unprovided SE of AUC of the validation data.*

1.19; 95% CI, 0.76-1.62; Appendix Figure 4B) survival. The corresponding AUCs were 0.71 for 90-day and 0.77 for 1-year survival. In SORG-MLA, the pooled results using the four US cohorts also showed great discriminatory ability for 90-day (logit(AUC), 1.50; 95% CI, 1.33-1.68; Appendix Figure 4C) and 1-year (logit(AUC), 1.72; 95% CI, 1.33-2.11; Appendix Figure 4D) survival. The corresponding AUCs were 0.82 for 90-day and 0.85 for 1-year survival. The sensitivity analysis excluding both developmental and internal validation cohort demonstrated similar results (Appendix Figure 5).

Both algorithms performed better in the USA than in Taiwan and the predictive power of SORG-CA showed less variability than SORG-MLA (Table 3 and Appendix Table 2). On one hand, the difference of logit(AUC)s in different countries were around 0.3 for SORG-CA and greater than 0.5 for SORG-MLA, and their 95% CIs did not overlap. On the other hand, multivariable regression suggested that only types of algorithms (i.e., SORG-CA vs SORG-MLA) and countries where the studies were conducted (i.e., the US vs Taiwan) had significant influence on discriminatory power. As the coefficient was significantly greater than 0, we can conclude that both algorithms

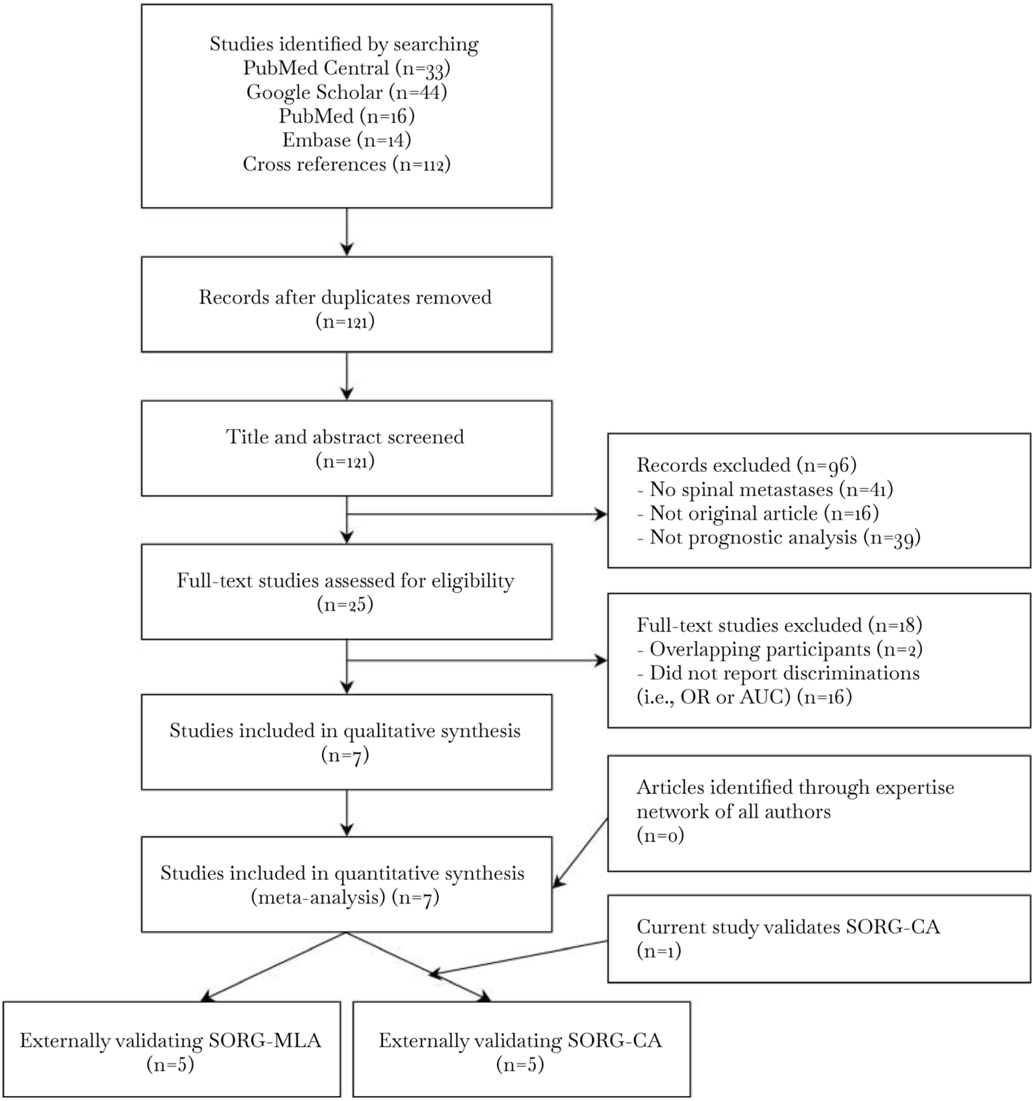

Studies identified by searching
PubMed Central (n=33)
Google Scholar (n=44)
PubMed (n=16)
Embase (n=14)
Cross references (n=112)
Records after duplicates removed
(n=121)
Title and abstract screened
(n=121)
Records excluded (n=96)
- No spinal metastases (n=41)
- Not original article (n=16)
- Not prognostic analysis (n=39)
Full-text studies assessed for eligibility
(n=25)
Full-text studies excluded (n=18)
- Overlapping participants (n=2)
- Did not report discriminations
(i.e., OR or AUC) (n=16)
Studies included in qualitative synthesis
(n=7)
Articles identified through expertise
network of all authors
(n=0)
Studies included in quantitative synthesis
(meta-analysis) (n=7)
Current study validates SORG-CA
(n=1)
Externally validating SORG-MLA
(n=5)
Externally validating SORG-CA
(n=5)

**Figure 1.** Flowchart of included studies. Two studies (Karhade et al, 2019 and Akash et al, 2021) validated both models and current studies provide two cohorts to validate SORG-CA, resulting in 5 cohorts for each model. OR=odds ratio; AUC=area under receiver operating characteristic curve; SORG=Skeletal Oncology Research Group; CA=classical algorithm; MLA=machine learning algorithm.

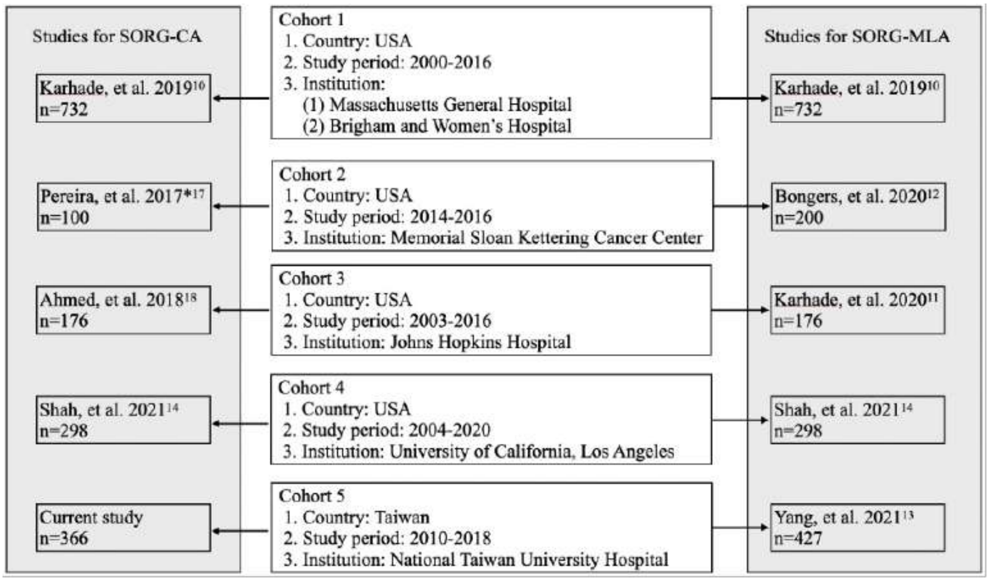

Studies for SORG-CA
Karhade, et al. 2019[10] n=732
Pereira, et al. 2017*[17] n=100
Ahmed, et al. 2018[18] n=176
Shah, et al. 2021[14] n=298
Current study n=366
Cohort 1
1. Country: USA
2. Study period: 2000-2016
3. Institution:
(1) Massachusetts General Hospital
(2) Brigham and Women's Hospital
Cohort 2
1. Country: USA
2. Study period: 2014-2016
3. Institution: Memorial Sloan Kettering Cancer Center
Cohort 3
1. Country: USA
2. Study period: 2003-2016
3. Institution: Johns Hopkins Hospital
Cohort 4
1. Country: USA
2. Study period: 2004-2020
3. Institution: University of California, Los Angeles
Cohort 5
1. Country: Taiwan
2. Study period: 2010-2018
3. Institution: National Taiwan University Hospital
Studies for SORG-MLA
Karhade, et al. 2019[10] n=732
Bongers, et al. 2020[12] n=200
Karhade, et al. 2020[11] n=176
Shah, et al. 2021[14] n=298
Yang, et al. 2021[13] n=427

**Figure 2.** Illustration of included cohorts per SORG algorithm. USA=United States of America; SORG=Skeletal Oncology Research Group; CA=classical algorithm; MLA=machine learning algorithm. *This study was based on cohort only in 2014.

**Table 3.** Summary and comparison of logit(AUC) by regional subgroup analysis

| Prediction algorithm | Pooled logit (AUC) with 95% CI | | Corresponding AUC | | Difference of logit (AUC) (SE) | P-value |
|---|---|---|---|---|---|---|
| | USA | Taiwan | USA | Taiwan | | |
| 90-day SORG-CA | 0.90 (0.56-1.24) | 0.58 (0.30-0.85) | 0.71 | 0.64 | 0.32 (0.04) | <0.001 |
| 90-day SORG-MLA | 1.50 (1.33-1.68) | 0.99 (0.71-1.28) | 0.82 | 0.73 | 0.51 (0.04) | <0.001 |
| 1-year SORG-CA | 1.19 (0.76-1.62) | 0.80 (0.53-1.07) | 0.77 | 0.69 | 0.39 (0.05) | 0.001 |
| 1-year SORG-MLA | 1.72 (1.33-2.11) | 1.05 (0.80-1.30) | 0.85 | 0.74 | 0.67 (0.04) | <0.001 |

*CI=confidence interval; AUC=area under receiver operating characteristic curves; SE=standard error of logit-transformed AUC; SORG=Skeletal Oncology Research Group; CA=classical algorithm; USA=United States of America.*

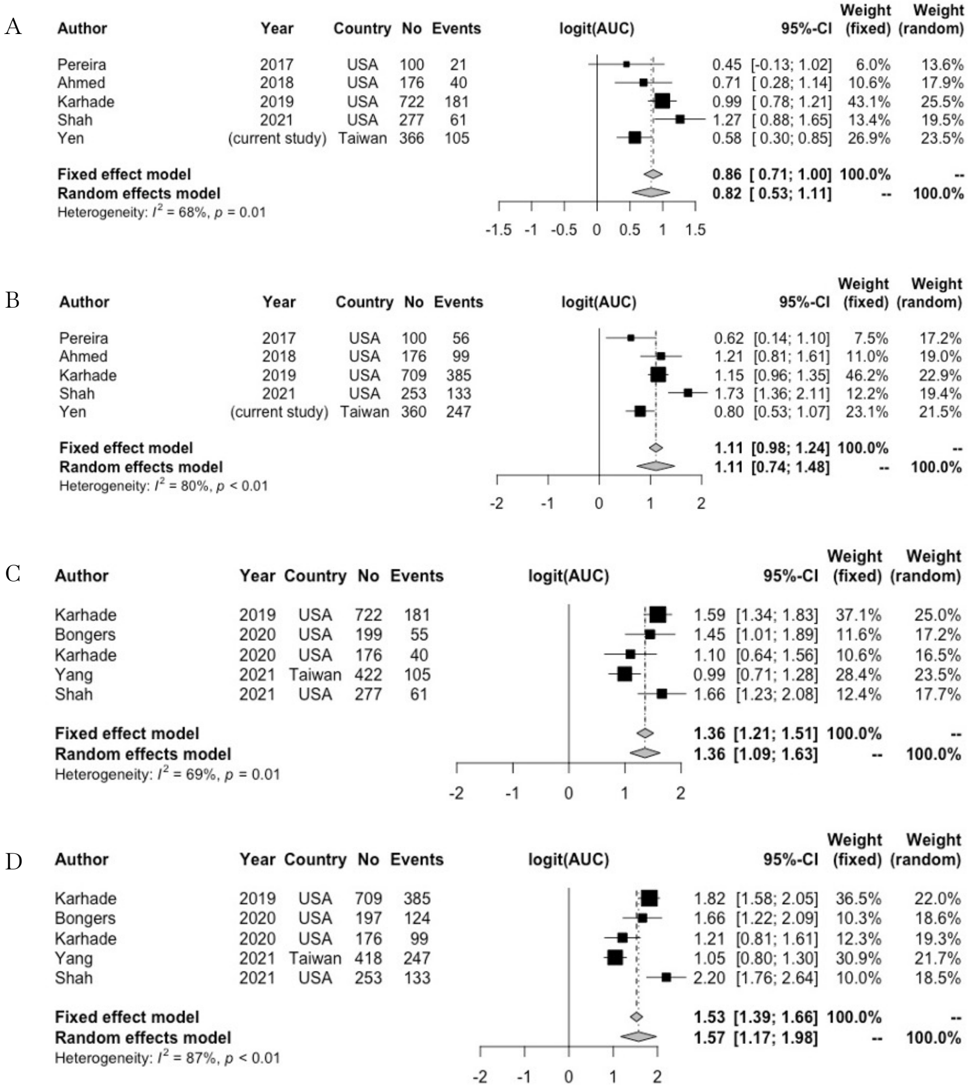


**Figure 3.** Forest plots and pooling odds ratio of (A) 90-day SORG-CA, (B) 1-year SORG-CA, (C) 90-day SORG-MLA, and (D) 1-year SORG-MLA. USA=United States of America; logit(AUC)=logit-transformed area under receiver operating characteristic curves; CI=confidence interval; SORG=Skeletal Oncology Research Group; CA=classical algorithm; MLA=machine learning algorithm. The patient number in the same study could vary since the survival of some patients could not be ascertained.

performed better in the US than in Taiwan. Of note, the discrimination was not varying significantly between the developmental and external validation cohorts. This might further substantiate that the algorithms generalize well and avoid overfitting to the developmental data.

### External Validation of SORG-CA

The current external validation cohort differed with the initial SORG-CA development cohort in the following eight variables: ECOG performance status, primary tumor histology, presence of visceral, brain, and multiple spine metastases, previous systemic therapy, preoperative WBC count, and 1-year mortality rate (Appendix Table 3). More than 98% of the Taiwanese cohort fall into the Chinese Han category defined by the US Census Bureau, indicating a racially distinct cohort from the initial development cohorts. The AUCs SORG-CA were 0.64 for 90-day (95% CI, 0.57-0.70) and 0.69 for 1-year (95% CI, 0.61-0.79) survival prediction (Table 2).

## DISCUSSION

This is the first study pooling the performance of the two SORG-CA and SORG-MLA survival algorithms in patients with spinal metastasis undergoing surgery and comparing them by region. SORG-CA and SORG-MLA both provide reliable survival estimations in five similar cohorts, and both generalized well in external validation cohorts. Although SORG-CA is more convenient for clinical use, SORG-MLA showed better discrimination in terms of both 90-day and 1-year survival prediction. Their performance was both influenced by validating on a non-American, Taiwanese cohort where they had a better estimation to Americans' prognosis than Taiwanese's. However, the performance of SORG-CA was less affected by validation on the non-American cohort compared with SORG-ML.

This study has several limitations. First, the US cohorts comprised a small proportion of Asians.[3] Due to their privacy policy, we were only able to conduct a subgroup analysis by region instead of race. Second, we could only indirectly support the argument that region-specific BMI could be more useful than BMI itself since the internal mechanism of SORG-MLA was undisclosed and the pseudonymized data was unavailable.[50] Furthermore, many confounding factors, not only BMI, could lead to the observations. Some prognosticators could also vary between countries, such as primary tumor type (Appendix Figure 6).[50] Besides, factors not included in SORG-MLA might also be associated with the observations. As mentioned in previous studies, systemic therapy is more affordable in Taiwan than in the US.[4,19,50] Future studies could focus on more important prognosticators in SORG-MLA or factors outside of SORG-MLA to better the model.[21] Also, since the detailed data was not obtained, we could not pool the sensitivity, specificity, and a pooling decision curve analysis. Furthermore, since SORG-CA gives integer scores, instead of survival probability, none of the studies related to SORG-CA reported calibration results (i.e., calibration slopes or intercepts). Therefore, we could not pool and compare calibration results from each algorithm in this study. This stepwise approach proposed by Steyerberg to assess the performance of clinical prediction models would have provided a more comprehensive picture of the two algorithms.[43] This again emphasizes the need for open publication and international collaboration. Lastly, the number of included articles was small, and only one of them is outside the US. Such a single-center cohort could potentially lead to sampling bias. Despite these limitations, our review provides valuable insights in the performance of the two SORG survival algorithms and suggests that performance of prognostic models may vary depending on different region or races.

With the statistical, imaging, and therapeutic progress of metastatic disease, many scoring systems were developed in the last two decades, such as Revised Tokuhashi Score, Modified Bauer Score, and New England Spinal Metastasis Score.[16,24,46] We would have been interested to include these models in this meta-analysis to compare all the existing models. However, it was impossible since limited available external validations and the used datasets in this meta-analysis were not available to validate the other models. Therefore, we only performed a meta-analysis and compared SORG-CA with SORG-MLA because there were four cohorts available for both models. Although both displayed good discriminatory ability, SORG-MLA displayed better discrimination. On the other hand, SORG-CA is easier to use since the application of SORG-MLA is web-based and requires reliable access to internet connectivity. This restriction makes SORG-MLA likely not available in all settings. SORG-CA can be especially useful in regions where internet is not readily available.

Many confounding factors could lead to a worse survival estimation to Taiwanese, such as different disease severity, tumor characteristics, and operation philosophy. Yang et al. hypothesized that it could also result from different demographics such as baseline BMI levels. The hypothesis is consistent with the obesity paradox.[9,17,25] This paradox relates that the extent of obesity, usually measured by BMI, is inversely associated with patients' mortality rate. Americans tend to have a higher BMI than Taiwanese (BMI: 29 vs. < 24).[6,10,18,22,51] Specifically, Asian-Americans have the lowest BMI among different races (average BMI=25) and all other races have an BMI over 28.[15,22,38] Given the divergent baseline BMI levels, SORG-MLA indicate at risk of underestimating the survival of Asian populations, especially 1-year survival prediction (Appendix Figure 6A and 6B).

The two algorithms share similar prognosticators with only SORG-MLA containing the following additional parameters: BMI, American Spinal Injury Association (ASIA) impairment scale, and various laboratory values. Although SORG-MLA performed overall better, the discrimination of SORG-CA was less influenced than SORG-MLA by validating in the Taiwanese cohort. It could be a small piece of evidence to indirectly support Yang's hypothesis that region-specific BMI might be considered a better prognosticator than BMI itself. Furthermore, baseline BMI was divergent in the two countries. An Asian patient with a BMI of 25 (which is above average) might have a better prognosis than a Caucasian patient with the same BMI (which is below average in the US) because the latter might indicate frailty or sarcopenia.[23,26,52,53] On the other hand, ASIA impairment scale, ECOG score, or other laboratory values, such as international normalized ratio (INR), had uniform standards. Patients with the same ASIA impairment scale or ECOG score should had similar neurological function or performance status. In addition, other input parameters' baseline levels did not differ considerable in different regions or among cohorts.[1,28,29,41,45,50] A good prediction model like SORG-MLA should demonstrate consistent discriminatory ability in both subgroups of "patients with impaired ASIA scale (i.e., ASIA scale of A-D)" and "patients with normal ASIA scale". Therefore, we did not consider the different baselines of ECOG score or ASIA scale between the reported cohorts and our Taiwanese cohort was the main cause of the observed decline of SORG-MLA's discrimination in the latter. It is our opinion that SORG-MLA, and prognostic models in general, could be further optimized by taking region-specific variables such as BMI into account. Such optimization could potentially benefit the fast-growing Asian-American population[2] in the United states.

After analyzing our data, we conducted another literature review. In the search, we found few previous meta-analysis studies focusing on the impact of geographical or racial variance on the performance of survival prediction models using machine learning algorithms'. Our study demonstrated the discriminatory ability of some machine learning algorithms could vary in different regions. This could be due to different prevalence, diagnostic criteria, and/or testing policies of certain diseases in different parts of the world. Some reviews did report findings similar to ours but did not provide a quantitative analysis. Therefore, we argue that more meta-analysis studies for machine learning algorithms should be conducted to evaluate these models' generalizability, especially in racially or geographically distinct regions. We also believe the current meta-analysis indeed sheds some light to future directions in which analysis of machine learning algorithms can be improved.

# CONCLUSION

SORG-MLA seems to be the better algorithm compared with SORG-CA. However, SORG-CA is more user friendly as it applies less input variables, is easy to use, and does not require access to internet. Prediction models show great potential in supporting clinical decision making, but future models should address the region-specific concerns highlighted in this meta-analysis. Existing prognostication models may need to recalibrate and optimize by considering region-specific variables such as BMI. This optimization requires an increased effort of international collaboration so more patients across racially distinct regions can benefit from these promising prediction algorithms.

# REFERENCES

1. International Diabetes Federation, Diabetes Atlas. **Diabetes prevalence (% of population ages 20 to 79) - Country Ranking.** Available at: https://www.indexmundi.com/facts/indicators/SH.STA.DIAB.ZS/rankings. Accessed 7 January 2021.

2. United States Census Bureau. **"Asian/Pacific American Heritage Month: May 2011".** Facts for Features, https://www.census.gov/newsroom/releases/archives/facts_for_features_special_editions/cb11-ff06.html; 2011 Accessed 7 January 2021.

3. **United States Census Bureau. By decades.** Available at: https://www.census.gov/programs-surveys/decennial-census/decade.html. Accessed 5 December 2020.

4. Aguiar PN, Haaland B, Park W, et al. **Cost-effectiveness of osimertinib in the first-line treatment of patients with EGFR-mutated advanced non-small cell lung cancer.** *JAMA Oncol* 4:1080-1084, 2018

5. Ahmed AK, Goodwin CR, Heravi A, et al: **Predicting survival for metastatic spine disease: a comparison of nine scoring systems.** *Spine J* 18:1804-1814, 2018

6. Alkhatib AL, Kreniske J, Zifodya JS, et al: **BMI is associated with coronavirus disease 2019 intensive care unit admission in african americans.** *Obesity (Silver Spring)* 28:1798-1801, 2020

7. Austin PC, Steyerberg EW. **Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable.** *BMC Med Res Methodol* 12:82, 2012

8. Bongers MER, Karhade AV, Villavieja J, et al. **Does the SORG algorithm generalize to a contemporary cohort of patients with spinal metastases on external validation?** *Spine J* 20:1646-1652, 2020

9. Caan BJ, Meyerhardt JA, Kroenke CH, et al: **Explaining the obesity paradox: the association between body composition and colorectal cancer survival (C-SCANS Study).** *Cancer Epidemiol Biomarkers Prev* 26:1008-1015, 2017

10. Cho D, Milbury K, McNeill LH. **Stress and cancer-related lifestyle factors among African American heterosexual couples.** *PLoS One 1*5:e0232577, 2020

11. Christodoulou E, Ma J, Collins GS, et al. **A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models.** *J Clin Epidemiol* 110:12-22, 2019

12. Cuschieri S. **The STROBE guidelines.** *Saudi J Anaesth* 13:S31-S34, 2019

13. Dea N, Versteeg AL, Sahgal A, et al. **Metastatic spine disease: should patients with short life expectancy be denied surgical care? An international retrospective cohort study.** *Neurosurgery* 87:303-311, 2020

14. Debray TP, Damen JA, Snell KI, et al. **A guide to systematic review and meta-analysis of prediction model performance.** *BMJ* 356:i6460, 2017

15. Fryar CD, Kruszon-Moran D, Gu Q, et al. **Mean body weight, height, waist circumference, and body mass index among adults: United States, 1999-2000 through 2015-2016.** *Natl Health Stat Report*:1-16, 2018

16. Ghori AK, Leonard DA, Schoenfeld AJ, et al. **Modeling 1-year survival after surgery on the metastatic spine.** *Spine J* 15:2345-2350, 2015

17. Hainer V, Aldhoon-Hainerova I. **Obesity paradox does exist.** *Diabetes Care* 36 Suppl 2:S276-281, 2013

18. Hsieh TH, Lee JJ, Yu EW, et al. **Association between obesity and education level among the elderly in Taipei, Taiwan between 2013 and 2015: a cross-sectional study.** *Sci Rep* 10:20285, 2020

19. Hsu JC, Wei CF, Yang SC. **Effects of removing reimbursement restrictions on targeted therapy accessibility**

for non-small cell lung cancer treatment in Taiwan: an interrupted time series study. *BMJ Open* 9:e022293, 2019

20. Karhade AV, Ahmed AK, Pennington Z, et al. External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease. *Spine J* 20:14-21, 2020

21. Karhade AV, Thio Q, Ogink PT, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. *Neurosurgery* 85:E671-E681, 2019

22. Lebel A, Kestens Y, Clary C, et al. Geographic variability in the association between socioeconomic status and BMI in the USA and Canada. *PLoS One* 9:e99158, 2014

23. Lee Y, Kim J, Han ES, et al. Frailty and body mass index as predictors of 3-year mortality in older adults living in the community. *Gerontology* 60:475-482, 2014

24. Leithner A, Radl R, Gruber G, et al. Predictive value of seven preoperative prognostic scoring systems for spinal metastases. *Eur Spine J* 17:1488-1495, 2008

25. Lennon H, Sperrin M, Badrick E, et al. The obesity paradox in cancer: a review. *Curr Oncol Rep* 18:56, 2016

26. Liao Q, Zheng Z, Xiu S, et al. Waist circumference is a better predictor of risk for frailty than BMI in the community-dwelling elderly in Beijing. *Aging Clin Exp Res* 30:1319-1325, 2018

27. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 339:b2700, 2009

28. Lutsey PL, Cushman M, Steffen LM, et al. Plasma hemostatic factors and endothelial markers in four racial/ethnic groups: the MESA study. *J Thromb Haemost* 4:2629-2635, 2006

29. Mills KT, Bundy JD, Kelly TN, et al. Global disparities of hypertension prevalence and control: a systematic analysis of population-based studies from 90 countries. *Circulation* 134:441-450, 2016

30. Morrison A, Polisena J, Husereau D, et al. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. *Int J Technol Assess Health Care* 28:138-144, 2012

31. Nater A, Tetreault LA, Kopjar B, et al. Predictive factors of survival in a surgical series of metastatic epidural spinal cord compression and complete external validation of 8 multivariate models of survival in a prospective North American multicenter study. *Cancer* 124:3536-3550, 2018

32. Nussbaumer-Streit B, Klerings I, Dobrescu AI, et al. Excluding non-English publications from evidence-syntheses did not change conclusions: a meta-epidemiological study. *J Clin Epidemiol* 118:42-54, 2020

33. Patchell RA, Tibbs PA, Regine WF, et al. Direct decompressive surgical resection in the treatment of spinal cord compression caused by metastatic cancer: a randomised trial. *Lancet* 366:643-648, 2005

34. Paulino Pereira NR, Janssen SJ, van Dijk E, et al. Development of a prognostic survival algorithm for patients with metastatic spine disease. *J Bone Joint Surg Am* 98:1767-1776, 2016

35. Paulino Pereira NR, McLaughlin L, Janssen SJ, et al. The SORG nomogram accurately predicts 3- and 12-months survival for operable spine metastatic disease: External validation. J *Surg Oncol* 115:1019-1027, 2017

36. Prasad D, Schiff D. Malignant spinal-cord compression. *Lancet Oncol* 6:15-24, 2005

37. Quraishi NA, Manoharan SR, Arealis G, et al. Accuracy of the revised Tokuhashi score in predicting survival in patients with metastatic spinal cord compression (MSCC). *Eur Spine J* 22 Suppl 1:S21-26, 2013

38. Reynolds K, He J. Epidemiology of the metabolic syndrome. *Am J Med Sci* 330:273-279, 2005

39. Shah AA, Karhade AV, Park HY, et al. Updated external validation of the SORG machine learning algorithms

for prediction of ninety-day and one-year mortality after surgery for spinal metastasis. *Spine J*, 2021:31;9430

40. Shaw B, Mansfield FL, Borges L. One-stage posterolateral decompression and stabilization for primary and metastatic vertebral tumors in the thoracic and lumbar spine. *J Neurosurg* 70:405-410, 1989

41. Singhal D, Smorodinsky E, Guo L. Differences in coagulation among Asians and Caucasians and the implication for reconstructive microsurgery. *J Reconstr Microsurg* 27:57-62, 2011

42. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28:112-118, 2012

43. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 35:1925-1931, 2014

44. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 283:2008-2012, 2000

45. Thrift AG, Thayabaranathan T, Howard G, et al. Global stroke statistics. *Int J Stroke* 12:13-32, 2017

46. Tokuhashi Y, Matsuzaki H, Oda H, et al. A revised scoring system for preoperative evaluation of metastatic spine tumor prognosis. *Spine (Phila Pa 1976)* 30:2186-2191, 2005

47. Ulmar B, Naumann U, Catalkaya S, et al. Prognosis scores of Tokuhashi and Tomita for patients with spinal metastases of renal cancer. *Ann Surg Oncol* 14:998-1004, 2007

48. Walter SD, Sinuff T. Studies reporting ROC curves of diagnostic and prediction data can be incorporated into meta-analyses using corresponding odds ratios. *J Clin Epidemiol* 60:530-534, 2007

49. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 170:51-58, 2019

50. Yang JJ, Chen CW, Fourman MS, et al. International external validation of the SORG machine learning algorithms for predicting 90-day and 1-year survival of patients with spine metastases using a Taiwanese cohort. *Spine J*, 2021:2;1529-9430

51. Yang TP, Chen HM, Hu CC, et al. Interaction of osteoarthritis and BMI on leptin promoter methylation in Taiwanese adults. *Int J Mol Sci* 21, 2019

52. Zakaria HM, Massie L, Basheer A, et al. Application of morphometrics as a predictor for survival in patients with prostate cancer metastasis to the spine. *World Neurosurg* 114:e913-e919, 2018

53. Zakaria HM, Wilkinson BM, Pennington Z, et al. Sarcopenia as a prognostic factor for 90-day and overall mortality in patients undergoing spine surgery for metastatic tumors: a multicenter retrospective cohort study. *Neurosurgery* 87:1025-1036, 2020

54. Zoccali C, Skoch J, Walter CM, et al. The Tokuhashi score: effectiveness and pitfalls. *Eur Spine J* 25:673-678, 2016

# SUPPLEMENTAL MATERIAL TO CHAPTER 15

**Appendix table 1.** Different searching terms used in different databases.

**Appendix table 2.** Results of subgroup analysis by meta-regression.

**Appendix table 3.** Baseline characteristics of external validation cohort (Taiwan) and development dataset (SORG-CA).

**Appendix figure 1.** PROBAST results of (A) included studies for SORG-CA (n=5), and (B) included studies for SORG-MLA (n=5). PROBAST, prediction model risk of bias assessment tool

**Appendix figure 2.** Flow of enrolling patients in this study.

**Appendix figure 3.** Forest plots and pooling odds ratio, which exclude the developmental and internal validation cohort, of (A) 90-day SORG-CA, (B) 1-year SORG-CA, (C) 90-day SORG-MLA, and (D) 1-year SORG-MLA. USA=United States of America; logit(AUC)=logit-transformed area under receiver operating characteristic curves; CI=confidence interval; SORG=Skeletal Oncology Research Group; CA=classical algorithm; MLA=machine learning algorithm. The patient number in the same study could vary since the survival of some patients could not be ascertained.

**Appendix figure 4.** Forest plots and pooling odds ratio of regional subgroup analysis, which exclude the developmental and internal validation cohort, of (A) 90-day SORG-CA, (B) 1-year SORG-CA, (C) 90-day SORG-MLA, and (D) 1-year SORG-MLA. USA=United States of America; logit(AUC)=logit-transformed area under receiver operating characteristic curves; CI=confidence interval; SORG=Skeletal Oncology Research Group; CA=classical algorithm; MLA=machine learning algorithm. The patient number in the same study could vary since the survival of some patients could not be ascertained.

**Appendix figure 5.** Forest plots and pooling odds ratio of regional subgroup analysis of (A) 90-day SORG-CA, (B) 1-year SORG-CA, (C) 90-day SORG-MLA, and (D) 1-year SORG-MLA. USA=United States of America; logit(AUC)=logit-transformed area under receiver operating characteristic curves; CI = confidence interval; SORG = Skeletal Oncology Research Group; CA = classical algorithm; MLA = machine learning algorithm. The patient number in the same study could vary since the survival of some patients could not be ascertained.

**Appendix figure 6.** Global variable importance for prediction of 90-day (left) and 1-year survival (right).

Supplemental material can be consulted online per the website of the journal and/or publisher.

CHAPTER

16

# NATURAL LANGUAGE PROCESSING FOR AUTOMATED QUANTIFICATION OF BONE METASTASES REPORTED IN FREE-TEXT BONE SCINTIGRAPH REPORTS

Olivier Q. Groot, Michiel E.R. Bongers, Aditya V. Karhade, Neal D. Kapoor, Brian P. Fenn, Jason Kim, Jorrit-Jan Verlaan, Joseph H. Schwab

*Acta Oncologica 2020;59:1455-1460*

# ABSTRACT

### Background

The widespread use of electronic patient-generated health data has led to unprecedented opportunities for automated extraction of clinical features from free-text medical notes. However, processing this rich resource of data for clinical and research purposes, depends on labor-intensive and potentially error-prone manual review.

### Objectives

To develop a natural language processing (NLP) algorithm for binary classification (single metastasis versus two or more metastases) in bone scintigraphy reports of patients undergoing surgery for bone metastases.

### Design

Natural language processing

### Methods

Bone scintigraphy reports of patients undergoing surgery for bone metastases were labeled each by three independent reviewers using a binary classification (single metastasis versus two or more metastases) to establish a ground truth. A stratified 80:20 split was used to develop and test an extreme gradient boosting supervised machine learning NLP algorithm.

### Results

A total of 704 free-text bone scintigraphy reports from 704 patients were included in this study and 617 (88%) had multiple bone metastases. In the independent test set (n.141) not used for model development, the NLP algorithm achieved an 0.97 AUC-ROC (95% confidence interval [CI], 0.92–0.99) for classification of multiple bone metastases and an 0.99 AUC-PRC (95% CI, 0.99–0.99). At a threshold of 0.90, NLP algorithm correctly identified multiple bone metastases in 117 of the 124 who had multiple bone metastases in the testing cohort (sensitivity 0.94) and yielded 3 false positives (specificity 0.82). At the same threshold, the NLP algorithm had a positive predictive value of 0.97 and F1- score of 0.96.

### Conclusion

NLP has the potential to automate clinical data extraction from free text radiology notes in orthopedics, thereby optimizing the speed, accuracy, and consistency of clinical chart review. Pending external validation, the NLP algorithm developed in this study may be implemented to aid

researchers in tackling large amounts of data.

# INTRODUCTION

In medicine, electronic health record (EHR) data is increasing exponentially over time.[1] The majority of this data is unstructured text in clinical reports, impeding its utilization in clinical practice and research setting. Manually extracting clinical characteristics of interest from these medical documents remain inefficient and prone to error; therefore neglecting potential valuable information.[2,3] One of these characteristics is the number of bone metastases as the quantity of bone metastases is associated with adverse outcomes such as postoperative complications and survival in oncologic populations.[4–6] No diagnosis code or automated extraction tool is available to bypass error prone and time-consuming manual chart review.

Artificial intelligence (AI) has emerged as a powerful method to transform medical care.[7–9] Although many AI-based methods have emerged in orthopaedic healthcare with strong performance, analysis of free-text clinical notes remains challenging.[5] One approach to analyze the free-text of patients' medical records is the use of natural language processing (NLP), a subfield of AI that focuses on enabling computers to process human language.[10] However, to our knowledge, there are no NLP algorithms available for extracting meaningful clinical features from free-text radiology reports in the field of orthopaedic oncology.

The aim of this study was to develop an NLP algorithm for binary classification (single metastasis versus two or more metastases) in bone scintigraphy reports of patients undergoing surgery for bone metastases.

# METHODS

### Study Design and Setting

The TRIPOD guidelines were followed for the development of the algorithm reported in this study.[11] Institutional review board approval was granted for retrospective review of EHRs. The inclusion criteria for this study were: (1) aged 18 years or older; (2) surgical treatment for a bone metastatic lesion; (3) date of procedure between January 1st, 2002 and January 1st, 2017; (4) index surgery at one of our two affiliated tertiary care hospitals; and (4) free-text bone scintigraphy reports within 6 months prior to the first index surgery in our institution's EHR available for review. Metastatic lesions were accounted for in the axial or appendicular skeleton, and also included multiple myeloma and lymphoma.[12] We excluded patients with (1) revision procedures, defined as any subsequent procedure after the index surgery addressing the metastatic lesion; and (2) kyphoplasty or vertebroplasty only.

The selection criteria were based on previous published studies – in which "single versus multiple bone metastases" a meaningful clinical feature was - that composed the cohort from which this current study extracted the bone scintigraphy reports. All patients in the cohort had at least a single bone metastasis. If a patient had multiple preoperative bone scintigraphy reports, the free-text report closest to surgery with a maximum of 6 months was obtained. If a patient underwent multiple surgeries, we considered the first surgery for bone metastases as the index procedure.

EHRs of patients in our institutional database of metastatic bone tumor were reviewed.[13,14] We identified 1780 potentially eligible patients after screening the medical records, of which 1076 patients did not have a preoperative bone scintigraphy within 6 months. A total of 704 radiology reports from 704 patients were included in this study (Figure 1).

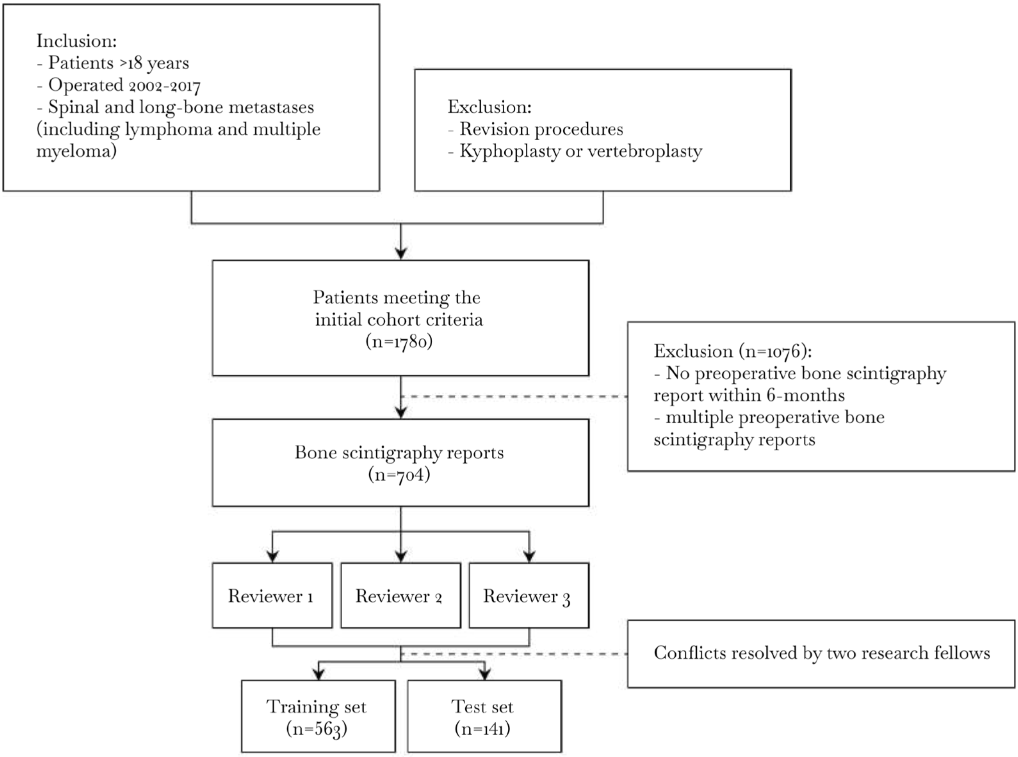

Inclusion:
- Patients >18 years
- Operated 2002-2017
- Spinal and long-bone metastases (including lymphoma and multiple myeloma)
Exclusion:
- Revision procedures
- Kyphoplasty or vertebroplasty
Patients meeting the initial cohort criteria (n=1780)
Exclusion (n=1076):
- No preoperative bone scintigraphy report within 6-months
- multiple preoperative bone scintigraphy reports
Bone scintigraphy reports (n=704)
Reviewer 1
Reviewer 2
Reviewer 3
Conflicts resolved by two research fellows
Training set (n=563)
Test set (n=141)

**Figure 1.** Flow diagram depicting the NLP selection and human interpretation. Training and test set split up in 80:20%.

### Ground Truth

The primary outcome was defined as single versus multiple bone metastases. This was manually annotated from free-text bone scintigraphy reports using a binary classification (single metastasis versus two or more metastases). The 704 selected reports were manually reviewed by three independent research coordinators (NK, BPF, JK). Each reviewer was blinded to the labels generated by the other reviewers. No additional clinical information was provided beside the free-text bone scintigraphy reports. Conflicts between the three reviewers were resolved by final research fellows (OQG, MERB) to establish a ground truth. The accuracy for the three reviewers was calculated with the Cohen's kappa as an interrater reliability estimate.

### Statistical Analysis

Prior analysis, the raw text notes required generic and approach-specific preprocessing steps. First, free-text reports were preprocessed in the following two ways: (1) cleaned from redundant or duplicate information (e.g., white spaces between paragraphs, time, date), line breaks, and stop words (e.g., "and", "for", "the"); and (2) stemming which reduces words into a common base or root (e.g., "increased" and "uptake" converted to "increas" and "uptak", respectively) (see Appendix 1). This transformed the raw text into the most parsimonious representation of the lexical meaning in a text note. Second, the bag-of-words representation method was applied to describe the relative frequency of words within a free-text. In this method, a matrix is created with rows for all free-text notes and columns for words (tokens) in the bone scintigraphy notes that correspond with the occurrence and frequency of words in the scintigraphy notes. Third, the term frequency-inverse document frequency (TF-IDF) was used to adjust for common and very rare words. This method reflects how important a word is to a document and measures the number of times that words appear in each document relative to the frequency of these words across all documents. The bag-of-words and TF-IDF were used as final input for the algorithm.

A stratified 80:20 split of the total dataset of 704 patients was done to create a training set (n=563) and independent test set (n=141). An extreme gradient boosting (XGBoost) machine learning algorithm was developed on the training set to detect multiple bone metastases.[15] The final model was evaluated on the independent test set, which was not used in developing the NLP model. The output of the NLP model is binary classification (single vs multiple bone metastases). We used the following metrics to assess the model performance: (1) discrimination [area under the receiver operating curve (AUC), precision-recall curve (PRC), area under the precision-recall curve, sensitivity (recall), specificity, negative-predictive value (NPV), positive predictive value (PPV), F1-score, negative likelihood ratio (LLR-), positive likelihood ratio (LLR+)]; (2) calibration (calibration slope and intercept); and (3) overall performance (Brier score).[16] The Brier score ranges from 0 (perfect prediction) to 1 (worst prediction). For correct interpretation of the Brier score a comparison should be performed with the null-model Brier score, which assigns a predicted probability equal to the observed prevalence of the outcome to each patient – in this study the prevalence of multiple bone metastases in the dataset. A Brier score lower than the null model Brier score indicates greater performance of the algorithm (Appendix 2).

Local explanations were provided to enable the ability to highlight individuals words used by the algorithm to determine single versus multiple bone metastases in individual free-text scintigraphy reports.[17] This figure will show features in green that increased the estimation of the likelihood of multiple metastases whereas the features in red are those that decreased the estimation of the likelihood of single metastases. Anaconda Distribution (Anaconda, Inc., Austin, Texas), Python (Python Software Foundation, Wilmington, Delaware), R version (The R Foundation, Vienna, Austria), and RStudio (RStudio, Boston, Massachusetts) were used for data analysis.

## RESULTS

A total of 704 free-text bone scintigraphy reports from 704 patients were included in this study and 617 (88%) had multiple bone metastases. The patients had a mean age of 62 (standard deviation of 12) and 374 (53%) were female. The interrater reliability was adequate; the three reviewers generally agreed with each other (kappa=0.8). In the independent test set (n=141) not used for model development, the NLP algorithm achieved AUC-ROC of 0.97 (Figure 2A), AUC-PRC of 0.99 (Figure 2B), calibration intercept of -0.41, and calibration slope of 0.73 for classification of single versus multiple bone metastases (Table 1). The Brier score for multiple bone metastases was 0.05 compared to the null model Brier score (score for algorithm that estimates a probability equal to the population prevalence of multiple metastases for every patient) of 0.011.

At a threshold of 0.10 and 0.90, the algorithm achieved a F1-score of 0.96 and 0.96, sensitivity of 0.99 and 0.94, specificity of 0.41 and 0.82, NPV of 0.88 and 0.67, and PPV of 0.92 and 0.97, respectively (Table 2). The algorithm, at the thresholds of 0.10 and 0.90, correctly classified the presence of multiple bone metastases in 123 and 117 reports (true positives) of the 124 who had multiple bone metastases in the testing cohort (sensitivity 0.99 and 0.94, respectively) and yielded 10 and 3 false positives (specificity 0.41 and 0.82, respectively). Local explanation of an actual free-text report demonstrated the specific words that drive towards (green) and against (red) classifying this report as a multiple bone metastasis (Figure 3); the algorithm used words such as "increas," "fractur", and "active" in the note to detect the occurrence of multiple bone metastases.

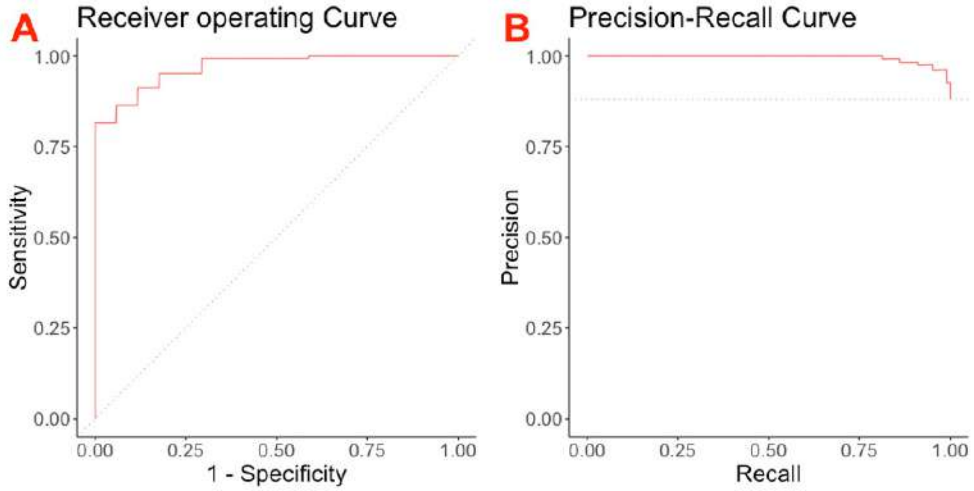
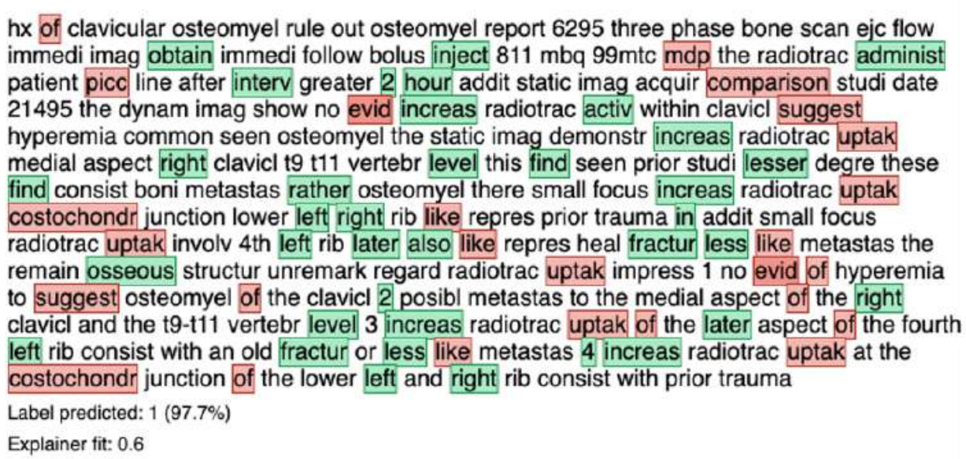


**Figure 2.** (A) Receiver operating curve and (B) Precision-Recall curves of NLP algorithm for multiple bone metastases in the independent testing set, n=141.



hx of clavicular osteomyel rule out osteomyel report 6295 three phase bone scan ejc flow
immedi imag obtain immedi follow bolus inject 811 mbq 99mtc mdp the radiotrac administ
patient picc line after interv greater 2 hour addit static imag acquir comparison studi date
21495 the dynam imag show no evid increas radiotrac activ within clavicl suggest
hyperemia common seen osteomyel the static imag demonstr increas radiotrac uptak
medial aspect right clavicl t9 t11 vertebr level this find seen prior studi lesser degre these
find consist boni metastas rather osteomyel there small focus increas radiotrac uptak
costochondr junction lower left right rib like repres prior trauma in addit small focus
radiotrac uptak involv 4th left rib later also like repres heal fractur less like metastas the
remain osseous structur unremark regard radiotrac uptak impress 1 no evid of hyperemia
to suggest osteomyel of the clavicl 2 posibl metastas to the medial aspect of the right
clavicl and the t9-t11 vertebr level 3 increas radiotrac uptak of the later aspect of the fourth
left rib consist with an old fractur or less like metastas 4 increas radiotrac uptak at the
costochondr junction of the lower left and right rib consist with prior trauma

Label predicted: 1 (97.7%)

Explainer fit: 0.6

**Figure 3.** Example of local explanation at the individual patient-level explanation for multiple bone metastases. By color-coding the algorithm visualizes which words influence the prediction positively (green) or negatively (red) toward the outcome, in this case the presence of multiple bone metastases. In addition, the algorithm provides a prediction percentage, and depending on the chosen threshold by the user, the algorithm generates a labeling of the outcome (depicted at the bottom).

**Table 1.** Overall performance of NLP algorithm for multiple bone metastases in the independent testing set (n=141)

| *Performance measures* | *NLP algorithm (95% CI)* |
|---|---|
| AUC-ROC | 0.97 (0.92-0.99) |
| AUC-PRC | 0.995 (0.986-0.999) |
| Brier | 0.05 (0.02-0.08) |
| Calibration intercept | −0.41 (−1.42-0.60) |
| Calibration slope | 0.73 (0.43-1.02) |
| Null model Brier score=0.11 | |

*AUC-PRC=area under the precision-recall curve; AUC-ROC=area under the receiver operating curve; NLP=natural language processing; CI=confidence interval*

**Table 2.** Performance of NLP algorithm at various thresholds for multiple bone metastases in the independent testing set (n=141)

| *Performance measures* | *NLP algorithm (95% CI)* | | |
|---|---|---|---|
| | *Threshold=0.90* | *Threshold=0.50* | *Threshold=0.10* |
| Sensitivity | 0.94 (0.89-0.98) | 0.98 (0.93-0.99) | 0.99 (0.96-1.00) |
| Specificity | 0.82 (0.57-0.96) | 0.71 (0.44-0.90) | 0.41 (0.18-0.67) |
| Negative predictive value | 0.67 (0.43-0.85) | 0.80 (0.52-0.96) | 0.88 (0.47-1.00) |
| Positive predictive value | 0.97 (0.93-0.99) | 0.96 (0.91-0.99) | 0.92 (0.87-0.96) |
| F1-score | 0.96 (0.91-0.99) | 0.97 (0.92-0.99) | 0.96 (0.91-0.98) |
| LLR (+) | 5.35 (1.91-14.9) | 3.32 (1.59-6.93) | 1.69 (1.13-2.51) |
| LLR (-) | 0.07 (0.03-0.15) | 0.03 (0.01-0.11) | 0.02 (0.00-0.15) |

*LLR (-)=negative likelihood ratio; LLR (+)=positive likelihood ratio; NLP=natural language processing; CI=confidence intervals*

# DISCUSSION

Many clinical features have no procedural or diagnosis code, making them subject to error prone and labor-intensive manual chart review. The amount of bone metastases is a characteristic that lacks these codes but is associated with adverse outcomes such as postoperative complications and survival in oncologic populations.[4-6] NLP constitutes a subfield of AI which shows promising results in analyzing the free-text included in EHRs.[10,18,19] The goal of this study was to develop an NLP algorithm for the binary classification of single and multiple bone metastases in bone scintigraphy reports of patients undergoing surgery for bone metastases. Our NLP algorithm correctly classified the presence of multiple bone metastases in 117 of the 124 (sensitivity 0.94) who had multiple bone metastases in the testing cohort and yielded only 3 false positives (specificity 0.82). Pending external validation, the NLP algorithm developed in this study may be implemented to aid clinicians and researchers in tackling large amounts of data.

This study has limitations. First, this was a retrospective study with clinical notes from tertiary hospitals from one health-care system. Multi-institutional cohorts and prospective, temporal, and external validation of the NLP algorithm remains to be conducted to support generalizability of the study findings to other medical institutions. Nevertheless, this study provides a framework and supports an innovative approach for developing NLP models for automating the analysis of free-text radiology notes. Second, the ground truth for binary classification was manual review. Despite being labeled by three independent reviewers, human classification remains prone to error.[2,3] However, using human consensus in establishing the ground truth is a commonly used method in the absence of an absolute ground truth.[20] Third, the NLP model was designed to classify single and multiple metastases in only bone scintigraphy reports. We did not design algorithms that would differentiate specific anatomic locations in reports of differing radiologic modalities. Future studies should incorporate the performance of NLP in non-bone scintigraphy radiology reports to quantify possible bone metastases and focus on differentiating the anatomical locations of bone metastases. Fourth, local explanation of the NLP algorithm identified some features (such as "fracture" or "evid") that appear to be clinically irrelevant to the presence of the bone metastases. Fracture may be clinically relevant because patients who had a pathologic fracture are more likely to have disseminated/advanced disease with multiple bone metastases. Words/tokens like "evid" may represent the features of radiologist lexicon when delineating multiple metastases in our cohort but may represent overfitting to the available data such that the models are not transportable to new, independent data. Moreover, although over 50 radiologists contributed to this dataset from two different hospitals, all radiology reports were from one health-care system with potentially use of fixed phrases to express certain type of findings. The algorithm may make accurate predictions in this study sample but may not generalize to other datasets. This emphasizes the need for external validation of the study findings to support generalizability of the NLP algorithm to other medical institutions. Fifth, future research may include other machine learning-based NLP algorithms such as convolutional and recurrent neural networks that may improve the performance demonstrated here. Sixth, over half of the patients were excluded due to the two exclusion criteria from this current study design. Comparing baseline characteristics demonstrated several differences between the included and excluded groups (Appendix 3). However, these clinical differences are not relevant for this study since it has no implications on the study aim or the developed NLP model as the model does not consider clinical, demographic, diagnosis, or treatment characteristics. Nevertheless, we deem the limitations proportionate to the strength of this NLP study. This study provides a proof-of-concept of applying similar NLP techniques to extract clinical features without procedural or diagnosis codes. To our knowledge, this is the first NLP study assessing an NLP algorithm for extracting clinical features without medical codes from free-text bone scintigraphy reports in the field of orthopaedic oncology. By using thorough crosschecked manual labeling, this study provides valuable insights into the use of NLP in in orthopaedics and its future role in clinical and research setting.

The manual process of extracting clinical features from free-text can be time-consuming and labor-intensive, and can therefore produce variable results.[3] With the recent widespread use of electronic medical records, the use of automated data extraction is on the rise.[1] However, few studies used NLP to explore classification analysis of free-text radiology reports for patients with metastases as well as other malignancies. Senders et al. previously used NLP to quantify brain metastases in magnetic resonance imaging reports.[21] Similarly, their NLP model had a high AUC of 0.92 and accuracy of 82%. Other NLP studies analyzing non-orthopaedic oncologic radiology notes report comparable high AUCs ranging from 0.91 to 0.99.[22–28] In accordance with these studies, with a modest dataset (n=1000), an NLP algorithm can be developed that extracts clinical features from free-text radiology notes. Compared to prior studies, this study developed algorithms capable of providing both estimations for likelihood of multiple bone metastases as well as explanations at the population and individual report level for multiple bone metastases.

The acceptability of an NLP algorithm's error rate depends on the application. For example, if the intention in research is to accelerate the efficiency of manual review, higher false positive errors rates are less concerning. The "loss" would be a reduced efficiency by increasing the number of charts reviewed. In clinical practice different error rates and evaluation metrics are important. For instance, achieving an <15% error rate in medical concept classification corresponds with human agreement on the same task[29]; however, the error tolerance in daily practice might be lower, such as in misclassifying history of allergies or comorbidities. When developing a NLP algorithm, the tradeoffs between performance metrics have implications on potential biases and should be guided by the nature of the NLP task.[30]

We believe the NLP methods presented in this study may be useful in a range of orthopaedics areas. First, a robust NLP tool could support research by rapidly identifying specific patients or diseases based on radiographical, pathological, or clinical findings. For example, creating a cohort of patients' multiple bone metastases can propel research in understanding the impact of skeletal related events in this complicated patient population. In addition, the clinical feature "single versus multiple bone metastases" can be used in various studies as a risk factor for an outcome, as was the case for the studies that supplied the bone-scintigraphy reports.[13,14] This could substantially reduce reviewer burden and error rate. Second, incorporating these NLP algorithms in EHRs may benefit population-based surveillance efforts. Third, NLP algorithms can be tailored for specific study designs; for example, the NLP algorithm developed in this study can extract clinical features that do not have specific administrative procedural or diagnosis code, such as the outcome in this study. Fourth, NLP algorithms can be used to "screen" radiology reports for important information that may have been inadvertently missed by clinicians in daily practice. However, in view of the variability and complexity of used language in radiology reports, together with an imperfect NLP model, we believe that this NLP algorithm currently remains to be restricted for research purposes.

## CONCLUSION

The widespread use of electronic patient-generated health data has led to unprecedented opportunities for research purposes. AI-based NLP methods enable us to automate the transformation of these unstructured free-text to clinical features, thereby optimizing the speed, accuracy, and consistency of clinical chart review. This study provides an NLP algorithm that has the potential to automate clinical data extraction from radiology notes in orthopaedics. Pending external validation, the NLP algorithm developed in this study may be implemented to aid clinicians and researchers in tackling large amounts of data.

# REFERENCES

1. Peterson ED. **Machine learning, predictive analytics, and clinical practice: can the past inform the present?** *JAMA - J Am Med Assoc.* 2019;322(23):2283–2284.

2. Mi MY, Collins JE, Lerner V, et al. **Reliability of medical record abstraction by non-physicians for orthopedic research.** *BMC Musculoskelet Disord.* 2013;14:181.

3. Cruz CO, Meshberg EB, Shofer FS, et al. **Interrater reliability and accuracy of clinicians and trained research assistants performing prospective data collection in emergency department patients with potential acute coronary syndrome.** *Ann Emerg Med.* 2009;54(1):1–7.

4. Paulino Pereira NR, Ogink PT, Groot OQ, et al. **Complications and reoperations after surgery for 647 patients with spine metastatic disease.** *Spine J.* 2019;19(1):144–156.

5. Karhade AV, Thio QCBS, Ogink PT, et al. **Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation.** *Neurosurgery.* 2019:1;85:671-681

6. Janssen SJ, Kortlever JTP, Ready JE, et al. **Complications after surgical management of proximal femoral metastasis: A retrospective study of 417 patients.** *J Am Acad Orthop Surg.* 2016;24(7):483–494.

7. Obermeyer Z, Emanuel EJ. **Predicting the future-big data, machine learning, and clinical medicine.** *N Engl J Med.* 2016;375(13):1216–1219.

8. Rajkomar A, Dean J, Kohane I. **Machine learning in medicine.** *N Engl J Med.* 2019;380(14):1347–1358.

9. Wallis C. **How artificial intelligence will change medicine.** *Nature.* 2019;576(7787):S48.

10. Nadkarni PM, Ohno-Machado L, Chapman WW. **Natural language processing: An introduction.** *J Am Med Informatics Assoc.* 2011;18(5):544–551.

11. Collins GS, Reitsma JB, Altman DG, et al. **Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement.** *BMC Med.* 2015;13:1.

12. Nathan SS, Healey JH, Mellano D, et al. **Survival in patients operated on for pathologic fracture: Implications for end-of-life orthopedic care.** *J Clin Oncol.* 2005;23(25):6072–6082.

13. Groot OQ, Ogink PT, Paulino Pereira NR, et al. **High risk of symptomatic venous thromboembolism after surgery for spine metastatic bone lesions: a retrospective study.** *Clin Orthop Relat Res.* 2019;477(7):1674–1686.

14. Groot OQ, Ogink PT, Janssen SJ, et al. **High risk of venous thromboembolism after surgery for long bone metastases: A retrospective study of 682 patients.** *Clin Orthop Relat Res.* 2018;476(10).

15. Chen T, Guestrin C. **Xgboost: A scalable tree boosting system.** *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM; 2016:785–794.

16. Brier GW. **Verification of forecasts expressed in terms of probability.** *Mon Weather Rev.* 1950;78(1):1–3.

17. Ribeiro MT, Singh S, Guestrin C. **"Why should i trust you?: Explaining the predictions of any classifier."** *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM; 2016:1135–1144.

18. Liang H, Tsui BY, Ni H, et al. **Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence.** *Nat Med.* 2019;25(3):433–438.

19. Karhade AV, Bongers MER, Groot OQ, et al. **Natural language processing for automated detection of incidental durotomy.** *Spine J.* 2020:20(5);695-700

20. Valizadegan H, Nguyen Q, Hauskrecht M. Learning classification models from multiple experts. *J Biomed. Inform.* 2013;46(6):1125–1135.

21. Senders JT, Karhade AV, Cote DJ, et al. Natural language processing for automated quantification of brain metastases reported in free-text radiology reports. *JCO Clin Cancer Informatics.* 2019;3:1–9.

22. Chen P-H, Zafar H, Galperin-Aizenberg M, et al. Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. *J Digit Imaging.* 2018;31(2):178–184.

23. Bozkurt S, Gimenez F, Burnside ES, et al. Using automatically extracted information from mammography reports for decision-support. *J Biomed Inform.* 2016;62:224–231.

24. Ping XO, Tseng YJ, Chung Y, et al. Information extraction for tracking liver cancer patients' statuses: From mixture of clinical narrative report types. *Telemed e-Health.* 2013;19(9):704–710.

25. Carrell DS, Halgrim S, Tran DT, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: The case of breast cancer recurrence. *Am J Epidemiol.* 2014;179(6):749–758.

26. Sippo DA, Warden GI, Andriole KP, et al. Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing. *J Digit Imaging.* 2013;26(5):989–994.

27. Sevenster M, Bozeman J, Cowhy A, et al. Automatically pairing measured findings across narrative abdomen CT reports. *AMIA Symp.* 2013;2013:1262–71.

28. Glaser AP, Jordan BJ, Cohen J, et al. Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. *JCO Clin Cancer Informatics.* 2018;2(2):1–8.

29. Uzuner Ö, South BR, Shen S, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc.* 2011;18(5):552–6.

30. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet.* 2011;12(6):417–28.

# SUPPLEMENTAL MATERIAL TO CHAPTER 16

**Appendix 1.** Example of raw and processed text

**Appendix 2.** Performance metrics explained

**Appendix 3.** Baseline comparison between included (n=704) and excluded group (n=1076)

Supplemental material can be consulted online per the website of the journal and/or publisher.

# PART V

# Strengths and Limitations of Artificial Intelligence

CHAPTER

17

# MACHINE LEARNING PREDICTION MODELS IN ORTHOPAEDIC SURGERY: A SYSTEMATIC REVIEW IN TRANSPARENT REPORTING

Olivier Q. Groot, Paul T. Ogink, Amanda Lans, Peter K. Twining, Neal D. Kapoor, William DiGiovanni, Bas J.J. Bindels, Michiel E.R. Bongers, Jacobien H.F. Oosterhoff, Aditya V. Karhade, Fetullah C. Öner, Jorrit-Jan Verlaan, Joseph H. Schwab

*Journal of Orthopaedic Research 2021*

# ABSTRACT

### Background

Machine learning (ML) studies are becoming increasingly popular in orthopaedics but lack a critically appraisal of their adherence to peer-reviewed guidelines.

### Objectives

(1) Evaluate quality and transparent reporting of machine ML prediction models in orthopaedic surgery based on the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement;

(2) Assess the risk of bias with the Prediction model Risk Of Bias ASsessment Tool (PROBAST) guidelines.

### Design

Systematic review.

### Methods

A systematic review was performed to identify all ML prediction studies published in orthopaedic surgery through June 18th, 2020. Studies were included if they evaluated ML models for any prediction in an orthopaedic surgery outcome such as survival, patient reported outcomes measures (PROMs), or complications. Exclusion criteria were (1) non-ML techniques (such as multivariable regression analysis), (2) conference abstracts, (3) non-English studies, (4) lack of full-text, and (5) non-relevant study types such as animal studies, letters to the editors, and case-reports. Two reviewers independently extracted data and discrepancies were resolved by discussion with at least two additional reviewers present.

### Results

After screening 7138 studies, 59 studies met the study criteria and were included. Across all studies, the overall median completeness for the TRIPOD checklist was 53% (interquartile range 47%-60%). TRIPOD items that were reported in less than 10% of studies were abstract (3%), model-building procedures (3%), and model specifications (8%). TRIPOD items that were reported in more than 90% of studies were data source (100%), overall interpretation (98%), limitations (97%), and specifying the objective (95%). As assessed by PROBAST, the overall risk of bias was low in 44% (n=26), high in 41% (n=24), and unclear in 15% (n=9). High overall risk of bias was driven by incomplete reporting of performance measures, inadequate handling of missing data, and use of small datasets with not enough number of outcomes.

**Conclusion**

Although the number of ML studies in orthopaedic surgery is increasing rapidly, over 40% of the existing models are at high risk of bias. Furthermore, over half incompletely reported their methods and/or performance measures. Until these issues are adequately addressed to give patients and providers trust in ML models, a considerable gap remains between the development of ML prediction models and their implementation in orthopaedic practice.

# INTRODUCTION

Prediction models for orthopaedic surgical outcomes based on machine learning (ML) are rapidly emerging. Such models, if adequately reported, can guide treatment decision making, predict adverse outcomes, and streamline perioperative healthcare management. However, transparent and complete reporting is required to allow the reader to critically assess the presence of bias, facilitate study replication, and correctly interpret study results. Unfortunately, previous studies have suggested that prediction models demonstrate incomplete, untransparent reporting of items such as study design, patient selection, variable definitions and performance measures.[1,2] To our knowledge, there is no systematic review that has assessed the completeness of reporting for the currently available prognostic ML models in orthopaedic surgery.

The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement was published in 2015 to improve the quality of reporting of prediction models.[3,4] It provides a guideline for essential elements of prediction model studies. The statement is endorsed by over ten leading medical journals and has been cited thousands of times. The Prediction model Risk Of Bias ASsessment Tool (PROBAST) was developed to assess risk of bias in prediction models by the Cochrane Prognosis group in 2019, and has been successfully piloted.[5] Both the PROBAST and TRIPOD had yet to be published at the time several ML prediction models for orthopaedic surgical outcome were developed; nonetheless, we believe they can be used as benchmarks for measuring quality of reporting and bias even if the prediction models were published before their introduction.

In this systematic review, we (1) evaluate the quality and completeness of reporting of prediction model studies based on ML for prognosis of surgical outcomes in orthopaedics according to their adherence to the TRIPOD statement, and (2) assess the risk of bias with the PROBAST.

# METHODS

### Systematic Literature Search

Registration in the PROSPERO international prospective register of systematic reviews was performed prior to study initiation and can be found online (registration number CRD42020206522). The study is reported according to the 2009 PRISMA guidelines.[6] A systematic search, in collaboration with a medical professional librarian, of the available literature was performed in PubMed, Embase, and the Cochrane Library for studies published up to June 18th, 2020. Different domains of medical subject headings (MeSH) terms and keywords were combined with 'AND'. Two domains with all related words were included in our search: ML and all possible orthopaedic specialties (Appendix 1). Two reviewers (PTO, OQG) independently screened and assessed all eligible studies based on predefined criteria (Figure 1).

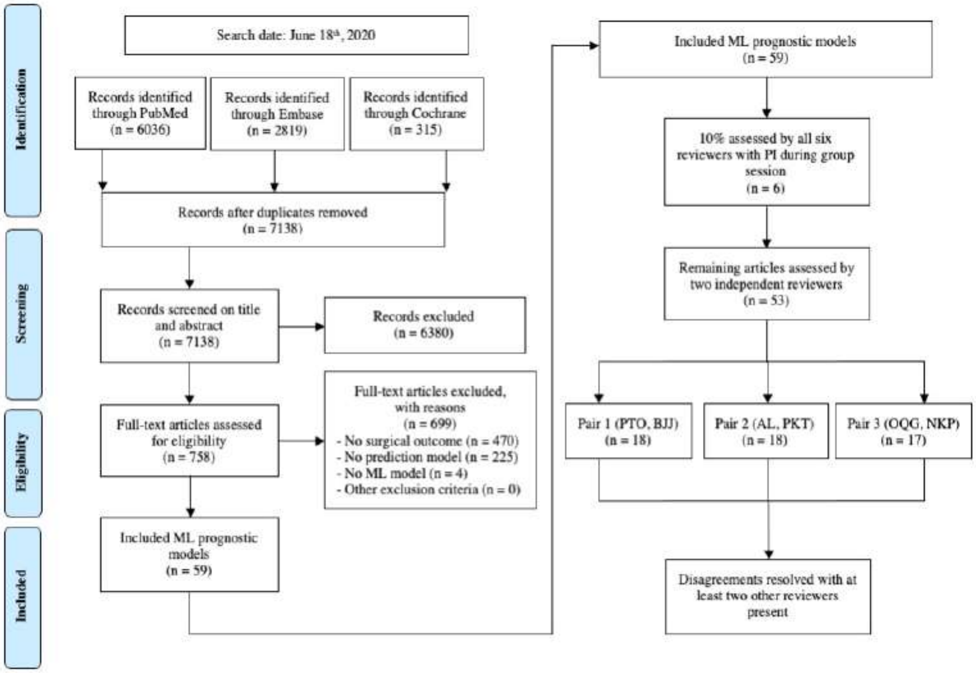

Identification
Screening
Eligibility
Included
Search date: June 18th, 2020
Records identified through PubMed (n = 6036)
Records identified through Embase (n = 2819)
Records identified through Cochrane (n = 315)
Records after duplicates removed (n = 7138)
Records screened on title and abstract (n = 7138)
Records excluded (n = 6380)
Full-text articles assessed for eligibility (n = 758)
Full-text articles excluded, with reasons (n = 699)
- No surgical outcome (n = 470)
- No prediction model (n = 225)
- No ML model (n = 4)
- Other exclusion criteria (n = 0)
Included ML prognostic models (n = 59)
Included ML prognostic models (n = 59)
10% assessed by all six reviewers with PI during group session (n = 6)
Remaining articles assessed by two independent reviewers (n = 53)
Pair 1 (PTO, BJJ) (n = 18)
Pair 2 (AL, PKT) (n = 18)
Pair 3 (OQG, NKP) (n = 17)
Disagreements resolved with at least two other reviewers present

**Figure 1.** PRISMA flowchart of study inclusions and exclusions. ML=machine learning; PI=principal investigator.

### Eligibility Criteria

Studies were included if they evaluated ML models for any prediction in an orthopaedic surgery outcome such as survival, patient reported outcomes measures (PROMs), or complications.

Exclusion criteria were (1) non-ML techniques (such as logistic or linear regression analysis), (2) conference abstracts, (3) non-English studies, (4) lack of full-text, and (5) non-relevant study types such as animal studies, letters to the editors, and case-reports. Orthopaedic specialties were defined as any operation for patients with musculoskeletal disorders.

### Data Extraction

Six reviewers (PTO, OQG, AL, PT, NDK, BBJ) independently assessed the first 10% of studies. All extracted data were then discussed during a group session with the principal investigator (PI) (JHS) to ensure quality and consistency. Any questions about discrepancies in the extracted data were resolved by the PI. After this quality training, the same six reviewers split up in pairs of two and each pair independently assessed the remaining 90% of studies which were evenly distributed among the three formed pairs. Each pair consisted of a research fellow with a medical doctor degree and a medical student. Disagreements within a pair were resolved during a consensus meeting with at least two other reviewers present. All six reviewers and the PI previously worked on and/or published ML prediction models in orthopaedic surgical outcomes.

For each included study, we extracted the following information: journal, prospective study design (yes/no), use of national or registry database (yes/no), size of total dataset, number of predictors used in final ML model, predicted outcome, mention of adherence to TRIPOD guideline in study (yes/no), access to ML algorithm (yes/no), TRIPOD items and PROBAST domains. The TRIPOD items and PROBAST domains are explained in more detail below.

The TRIPOD statement consists of 22 main items, of which two main items (12 and 17) refer to model updating or external validation studies, leaving 20 main items to be extracted for prognostic prediction modeling studies[4]. These main items were transformed into an adherence assessment form by the statement developers. Of the 20 main items, 11 had no subitems (1, 2, 8, 9, 11, 16, 18, 19, 20, 21, and 22), seven were divided into two subitems (e.g. 3a and 3b; 3, 4, 6, 7, 13, 14, and 15), and two into three subitems (e.g. 5a, 5b, 5c; 5 and 10). Four subitems (10c, 10e, 13c, and 19a) were, together with the two main items (12 and 17), not extracted because they did not refer to developmental studies (e.g. 10c "For validation, describe how the predictions were calculated"; Appendix 2). Hereafter, subitems and main items are defined under one nomenclature "items" (e.g. main item 3 consists of two items; 3a and 3b). In total, 29, 30, or 31 potential items could be assessed per study. This total number of items varied between 29 and 31 because some items could be scored with "not applicable" (e.g. 14b "if nothing on univariable analysis (in methods or results) is reported, score not applicable") and this was excluded when calculating the completeness of reporting. Also, some items could be scored with "referenced" (e.g. item 6a) Referenced was considered "completed" and included when calculating the completeness of reporting.

Each item may consist of multiple elements. Both elements must be scored "yes" for the item to be scored "completed." To calculate the completeness of reporting of TRIPOD items, the number of completely reported TRIPOD items was divided by the total number of TRIPOD items for that study. If a study reported on multiple prediction models (e.g. prediction model for 90-day and 1-year survival), we extracted data only on the best performing model.

PROBAST assesses the risk of bias in prognostic prediction model studies.[5] This tool consists of 20 signaling questions across four domains: participants selection (1), predictors (2), outcome (3), and analysis (4). Each domain is rated "low", "high", or "unclear" risk of bias. 'Unclear" indicates that the reported information is insufficient – no reliable judgement on low or high risk of bias can be made with the information provided. Participants selection (1) covers potential sources of bias in the origin of data and criteria for participant selection – are all patients included and excluded appropriately? Predictors (2) should include a list of all considered predictors, a clear definition and timing of measurement. An outcome (3) should include clear definitions and timing of measurements, and a description of the time interval between predictor assessment and outcome determination. Lasty, analysis (4) covers potential sources of bias related to inappropriate analysis methods or omission of key performance measures such as discrimination and calibration.

The ratings of the four domains resulted in an overall judgement about risk of bias. Low overall risk of bias was assigned if each domain scored low. High overall risk of bias was assigned if at least one domain was judged to be high risk of bias. Unclear overall risk of bias was noted if at least one domain was judged unclear and all other domains low. The four domains and the overall judgement were reported – not every signaling question.

### Statistical Analysis

Completeness of reporting of TRIPOD items and PROBAST domains were visualized by bar graphs. We used Microsoft Excel Version 19.11 (Microsoft Inc, Redmond, WA, USA) to extract and record data using standardized forms, Stata® 14.0 (StataCorp LP, College Station, TX, USA) for the statistical analyses, and Mendeley Desktop Version 1.19.4 (Mendeley Ltd, London, UK) as reference management software.

## RESULTS

The conducted search yielded 7,138 unique studies. Seven hundred and fifty-eight potential studies were selected by title and abstract screening, of which 59 remained after full-text screening (Appendix 3). Table 1 lists the study characteristics of the included study. The majority (83%; 49/59) was published after the launch of the TRIPOD statement (see Appendix 4). The 59 studies were published in 33 different medical journals of which three journals published 31% of all included studies (18/59). None of the studies were published in a journal that requested adherence to the TRIPOD guidelines in their instructions to authors.

### TRIPOD

Among all studies, the overall median completeness for the TRIPOD items was 53% (IQR 47%-60%; see Figure 2 and Appendix 5). Eight items were reported in over 75% of studies and seven items in less than 25% (Table 2). The abstract (2) and the model-building procedure (10b) were the most poorly reported items with only 3% (2/59). Source of data (4a) was reported in all studies (100%; 59/59).

**Table 1.** Characteristics of included studies (n=59)

| *Variables* | *median (IQR)* |
|---|---|
| Sample size | 4782 (616-23.264) |
| Predictors included in final model[a] | 10 (7-14) |
| | *% (n)* |
| Year of publication | |
| <2015 (prior to TRIPOD guideline) | 17 (10) |
| >2016 | 83 (49) |
| Number of publications per journal | |
| <5 publications per journal | 69 (41) |
| >5 publications per journal | 31 (18) |
| Prospective database | 3 (5) |
| National/Registry database[b] | 51 (30) |
| Mention of using TRIPOD | 20 (12) |
| Predicted outcome | |
| Complications | 24 (14) |
| PROM | 20 (12) |
| Mortality | 19 (11) |
| Health management | 19 (11) |
| Other | 19 (11) |

*TRIPOD=Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; ML=machine learning; PROM=Patient Reported Outcome Measure;*
*a The amount of predictors that were included in the final, best performing machine learning algorithm. In 14% (8/59) this could not be extracted from the study or was unclear.*
*b This includes databases such as Surveillance, Epidemiology, and End Results (SEER) or American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP).*

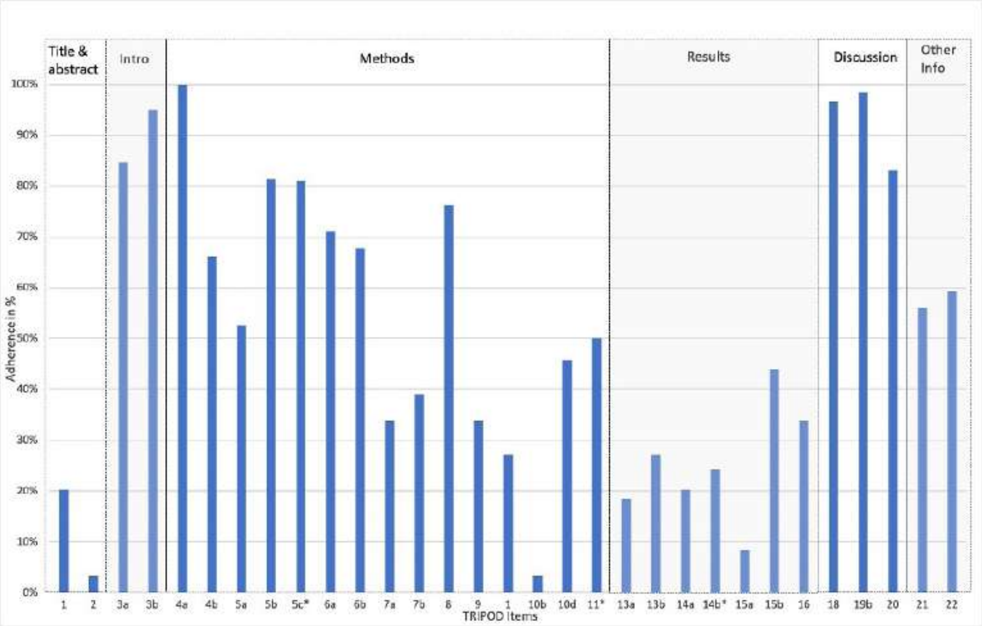


**Figure 2.** Overall adherence per TRIPOD item. *All items consisted of 59 datapoints, except for item 5c (58), item 11 (4) and item 14b (45) due to the "Not applicable" option.

The overall risk of bias was low in 44% (26/59), high in 41% (24/59), and unclear in 15% (9/59) of the studies (Figure 3.). The studies that rated highly for overall risk of bias were mainly rated this way due to bias in the analysis domain, (as opposed to the other three domains) incomplete reporting of performance measures, inadequate handling of missing data, or use of small datasets with low number of outcomes. Most notable was the lack of adequate reporting of performance measures such as calibration results, Brier scores, or decision-curves. Unclear risk of bias in the analysis domain was scored in 20% (12/59), mainly due to the lack of mention as to how continuous and categorical predictors were handled or how the handling of complexities in the data was reported (e.g. competing risk analysis).

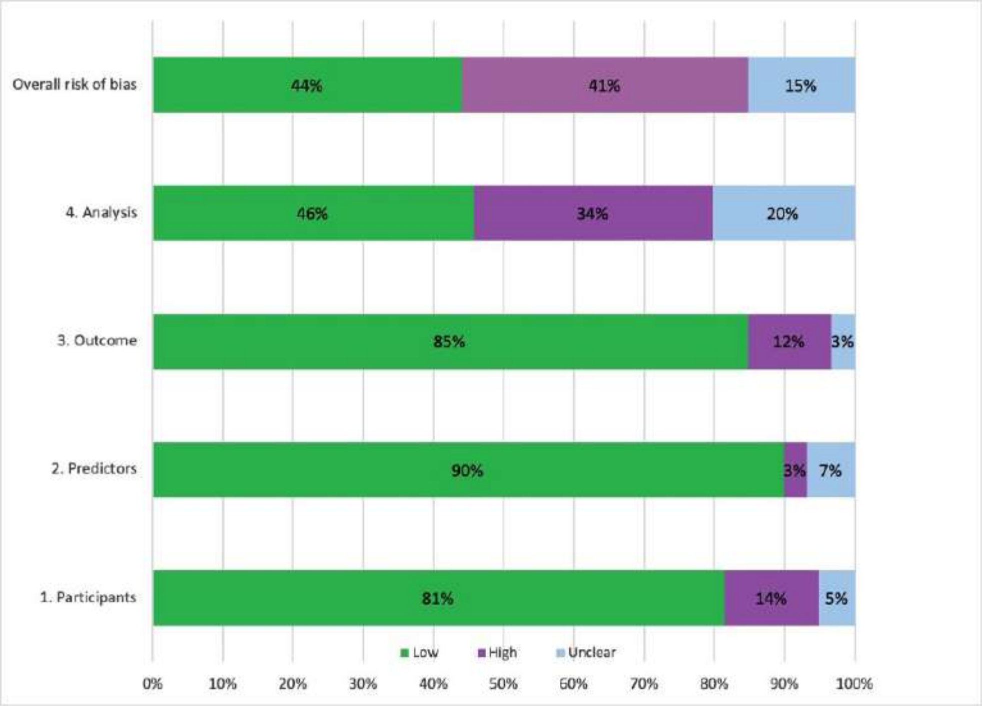


**Figure 3.** PROBAST results for all included studies (n=59).

**Table 2.** Individual TRIPOD items sorted by completeness of reporting over 75% and under 25%.

| *Complete reporting > 75%* | | | *Complete reporting < 25%* | | |
|---|---|---|---|---|---|
| *TRIPOD item* | *TRIPOD description* | *% (n)* | *TRIPOD item* | *TRIPOD description* | *% (n)* |
| 4a | Describe the study design or source of data (e.g., randomized trial, cohort, or registry data). | 100% (59) | 10b | Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation. | 3% (2) |
| 19b | Give an overall interpretation of the results considering objectives, limitations, results from similar studies and other relevant evidence. | 98% (58) | 2 | Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions. | 3% (2) |
| 18 | Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data). | 97% (57) | 15a | Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point). | 8% (5) |
| 3b | Specify the objectives, including whether the study describes the development of the model. | 95% (56) | 13a | Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful. | 19% (11) |
| 3a | Explain the medical context and rationale for developing the multivariable prediction model, including references to existing models. | 85% (50) | 14a | Specify the number of participants and outcome events in each analysis. | 20% (12) |
| 5b | Describe eligibility criteria for participants. | 83% (49) | 1 | Identify the study as developing a multivariable prediction model, the target population, and the outcome to be predicted. | 20% (12) |
| 5c* | Give details of treatments received, if relevant. | 81% (48) | 14b* | If done, report the unadjusted association between each candidate predictor and outcome. | 24% (11) |
| 8 | Explain how the study size was arrived at. | 76% (45) | | | |

*TRIPOD=Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.*
**All items consisted of 59 datapoints, except for 5c (58) and 14b (45) due to "Not applicable" option.*

# DISCUSSION

In this systematic review we aimed to assess the quality and transparency of reporting of currently published ML prediction models in surgical outcome in orthopaedics using the TRIPOD and PROBAST guidelines. The reporting of the study abstract had the worst adherence in existing models. According to the PROBAST, 41% of the studies displayed a high risk of bias, primarily due to risk of bias in the analysis domain. ML prediction models may support clinical decision making, but future studies should adhere to recognized methodological standards in order to develop ML prediction models of clinically significant value to healthcare professionals.

This review has several limitations. First, despite using a comprehensive search term in multiple online medical libraries, we may have missed some publications. However, we do not believe that these missed publications would have had a profound impact on the completeness of our reporting or on the final conclusions. Considering the large number of included studies, adding potentially missed studies would most likely not change our main conclusions that the overall adherence is poor. Second, TRIPOD guidelines were employed as a reporting benchmark. However, the relative importance of each item and what composes an acceptable score is up for debate. Third, a strict adherence to scoring was implemented on all elements of a TRIPOD item. For example, item 2 "Abstract" consists of 12 elements which all have to be fulfilled in order for item 2 to be marked as "completely reported". Also, authors as well as journal reviewers might have good reasons to exclude certain TRIPOD information. For example, one may not report regression coefficients in item 15 "model specifications" or provide "the potential clinical use of the model" in item 20 if they believe that their prediction model is not fit for clinical use. Nonetheless, we scored these items in this current study as "incomplete". This rigorous method of scoring is in line with the nature of the TRIPOD guideline and is deemed essential for consistent and transparent reporting of prediction models. In addition, most journals require a maximum word count or prescribe specific requirement. These restrictions could potentially prevent authors from including all 12 elements. Despite these limitations, this review provides the first comprehensive overview of completeness of transparent reporting for ML prediction models in orthopaedics. Illustrating poor reporting of TRIPOD items identifies current hurdles and may improve future transparent reporting.

The TRIPOD statement was published in 2015 to provide a framework for transparent reporting and quality of prediction models. Despite being published in 11 medical journals and being well-referenced 24% [12/49] of included studies published after the TRIPOD statement referenced TRIPOD. A possible explanation is the usual slow implementation of guidelines after publication.[7–12] Although the 11 medical journals are leading, high impact journals, none are orthopaedic specific journals so they may have been missed by the orthopaedic community. Another reason could be that authors of ML models have been dissuaded to adhere to TRIPOD doubting its applicability to their study. The explanatory documents of the TRIPOD statement focus on models based on regression techniques and several items do not fully pertain to ML, e.g. item 15a on regression coefficients. The authors of the TRIPOD statement recently acknowledged this drawback and have announced the development of a version specific to ML, TRIPOD-ML, similar to the CONSORT-AI extension.[13,14]

Alternative reasons for incomplete items are reviewers demanding different information than the items in TRIPOD, journal format and maximum word count limiting the number of items to mention, or researchers only using reporting guidelines near the end when writing up the manuscript. A study by Agha et al.[15] found considerable improvement in reporting was achieved after a surgical journal started mandating reporting guideline checklists to be included in the submission to the editor and reviewers. This could trigger researchers to include reporting guidelines like TRIPOD or ARRIVE (Animal Research: Reporting In Vivo Experiments)[16] in the early stages of study design instead of during manuscript writing, which according to Dewey et al. led to increased perceived value of the reporting guidelines.[17] However, adherence to TRIPOD is not a panacea. Logullo et al.[18] argue adherence to guidelines does not equal quality despite often being interpreted that way. For the TRIPOD statement it is important to stress the relative importance of each item as well as what constitutes a "good" score is debatable. For example, the omission of any calibration measure is arguably worse than incomplete reporting of the title. Nonetheless, in this relatively new research field it is a useful framework for standardization of reporting and researchers should strive to adhere to the TRIPOD statement.

According to the PROBAST assessment numerous studies were at high risk of bias. Predominantly, three area in the analysis domain were poorly scored. First, most models were built on databases with missing values, mostly due to use of national or registry databases such as NSQIP. Most often, predictors with incomplete data were excluded in the model building process. Both may lead to confounding or selection bias.[19,20] In other words, variables with a strong predictive accuracy may be missed or misinterpreted. This highlights the importance of preferably using prospective, complete datasets, and when missing data are present, processing them appropriately through techniques such as multiple imputation.[21]

A second issue is the incomplete reporting of performance measures. The vast majority of studies describe discrimination measures, predominantly area under the curve, while only a minority report calibration measure. Calibration is an essential element of describing the performance of ML models and its importance has extensively been discussed in earlier reviews.[22–24] The frequent omission of calibration renders assessment of performance virtually impossible and is in line with previous literature on prediction models.[2,25,26]

Finally, the small sample sizes with often small outcome numbers introduce risk of overfitting. Overfitting refers to including too many prognostic factors relative to the number of cases. This may improve the prediction performance in the dataset but reduces the generalizability outside the training dataset. While the use of national databases may circumvent the issue of small sample sizes, they have the disadvantage of oftentimes less granular data (e.g., lacking PROM scores), missing data, as highlighted earlier, and may lack important predictors such as laboratory values.[27]

Our findings lead to some careful recommendations for researchers developing ML prediction models. First, authors should mind all the necessary steps in model development and reporting, starting at the early stages of study design; the TRIPOD checklist can be a guiding tool to this end. Second, next to discrimination and calibration, model performance should always include a measure of clinical utility for decision-making. Decision-making analysis has been around for a significant amount of time, but has only recently started gaining popularity as a valuable tool in prediction models.[22,28] In short, decision-making analysis measures the net benefit of using the ML model prediction across the entire spectrum of predictions by weighing both the benefits for certain patients (true-positives) and the harm for other patients (false-positives). This is preferably assessed and visualized using decision curve analysis.[29]

Third, mere development of clinical prediction models is not the end goal, as they are eventually intended to be used in clinical practice. Prior to utilization by the medical community, extensive external validation is required to ensure robustness of the model outside the database used for development. However, less than half of the published studies offered means to calculate predictions through web calculators or in-study formulas, making external validation and individual predictions difficult.[30] Ideally, the algorithms are published online to facilitate sharing and collaboration.

# CONCLUSION

Prognostic surgical outcome models are rapidly entering the orthopaedic field to guide treatment decision making. This review indicates that numerous studies display poor reporting and are at high risk of bias. Future studies aimed at developing prognostic models should explicitly address the concerns raised, such as incomplete reporting of performance measures, inadequate handling of missing data, and not providing means to make individual predictions. Collaboration for sharing data and expertise is needed not just for developing more reliable prediction models, but also for validating current models. Methodological guidance such as the TRIPOD statement should be followed, for unreliable prediction models can cause more harm than benefit when guiding medical decision making.

# REFERENCES

1. Groot OQ, Bongers MER, Ogink PT, et al. Does artificial intelligence outperform natural intelligence in interpreting musculoskeletal radiological studies? A systematic review. *Clin Orthop Relat Res.* 2020;478:2751-2764

2. Wang W, Kiik M, Peek N, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLoS One.* 2020;15(6):e0234722.

3. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur Urol.* 2015;67(6):1142–1151.

4. Heus P, Damen JAAG, Pajouheshnia R, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. *BMC Med.* 2018;16(120).

5. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med.* 2019;170(1):51–58.

6. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev.* 2015;4(1):1.

7. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA.* 2001;285(15):1992–1995.

8. Korevaar DA, van Enst WA, Spijker R, et al. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. *Evid Based Med.* 2014;19(2):47–54.

9. Sekula P, Mallett S, Altman DG, et al. Did the reporting of prognostic studies of tumour markers improve since the introduction of REMARK guideline? A comparison of reporting in published articles. *PLoS One.* 2017;12(6):e0178531.

10. Smidt N, Rutjes AWS, van der Windt DAWM, et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? *Neurology.* 2006;67(5):792–797.

11. Turner L, Shamseer L, Altman DG, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database Syst Rev.* 2012;11(11):MR000030.

12. Chan A-W, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet (London, England).* 2005;365(9465):1159–1162.

13. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ.* 2020;370:m3164.

14. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet (London, England).* 2019;393(10181):1577–1579.

15. Agha RA, Fowler AJ, Limb C, et al. Impact of the mandatory implementation of reporting guidelines on reporting quality in a surgical journal: A before and after study. *Int J Surg.* 2016;30:169–172.

16. Percie du Sert N, Hurst V, Ahluwalia A, et al. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. *J Physiol.* 2020;598(18):3793–3801.

17. Dewey M, Levine D, Bossuyt PM, et al. Impact and perceived value of journal reporting guidelines among Radiology authors and reviewers. *Eur Radiol.* 2019;29(8):3986–3995.

18. Logullo P, MacCarthy A, Kirtley S, et al. Reporting guideline checklists are not quality evaluation forms: they

are guidance for writing. *Heal Sci reports*. 2020;3(2):e165.

19. Paxton C, Niculescu-Mizil A, Saria S. Developing predictive models using electronic medical records: challenges and pitfalls. *AMIA Symp*. 2013;2013:1109–1115.

20. Skelly AC, Dettori JR, Brodt ED. Assessing bias: the importance of considering confounding. *Evid Based Spine Care J*. 2012;3(1):9–12.

21. Li P, Stuart EA, Allison DB. Multiple imputation: a flexible tool for handling missing data. *JAMA*. 2015;314(18):1966–1967.

22. Karhade AV, Schwab JH. CORR synthesis: when should we be skeptical of clinical prediction models? *Clin Orthop Relat Res*. 2020;478:2722-2728

23. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138.

24. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928–935.

25. Hodgson A, Helmy N, Masri BA, et al. Comparative repeatability of guide-pin axis positioning in computer-assisted and manual femoral head resurfacing arthroplasty. *Proc Inst Mech Eng H*. 2007;221(7):713–724.

26. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ*. 2020;369:m1328.

27. Janssen DMC, van Kuijk SMJ, d'Aumerie BB, et al. External validation of a prediction model for surgical site infection after thoracolumbar spine surgery in a Western European cohort. *J Orthop Surg Res*. 2018;13(1):114.

28. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. 2006;26(6):565–574.

29. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925–1931.

30. Groot OQ, Bindels BJJ, Ogink PT, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop*. 2021:1–9.

# SUPPLEMENTAL MATERIAL TO CHAPTER 17

**Appendix 1.** Search syntaxes for the PubMed, Embase, and Cochrane Library on June 18th, 2020
**Appendix 2.** Completeness of reporting of individual TRIPOD items.
**Appendix 3.** List of included studies (n=59).
**Appendix 4.** Published studies per year (n=59).
**Appendix 5.** Overall TRIPOD adherence per study (n=59).

Supplemental material can be consulted online per the website of the journal and/or publisher.

CHAPTER

18

# AVAILABILITY AND REPORTING QUALITY OF EXTERNAL VALIDATIONS OF MACHINE LEARNING PREDICTION MODELS WITH ORTHOPEDIC SURGICAL OUTCOMES: A SYSTEMATIC REVIEW

Olivier Q. Groot, Bas J.J. Bindels, Paul T. Ogink, Neal D. Kapoor, Peter K. Twining, Austin K. Collins, Michiel E.R. Bongers, Amanda Lans, Jacobien H.F. Oosterhoff, Aditya V. Karhade, Jorrit-Jan Verlaan, Joseph H. Schwab

*Acta Orthopedica 2021;92:385-393*

# ABSTRACT

### Background

External validation of machine learning (ML) prediction models is an essential step before clinical application.

### Objectives

The study aims were to assess the (1) proportion, (2) performance and (3) transparent reporting of externally validated ML prediction models in orthopaedic surgery, using the Transparent Reporting for Individual Prognosis or Diagnosis (TRIPOD) guidelines.

### Design

Systematic review.

### Methods

A systematic search in PubMed, Embase, and the Cochrane Library was performed using synonyms for every orthopaedic specialty, ML, and external validation published up until November 17th, 2020. Inclusion criteria were external validation; prediction models based on ML; and orthopaedic surgical outcomes (defined as any outcome after musculoskeletal surgery). Exclusion criteria were non-ML prediction model (e.g., logistic regression); internal validation (e.g., cross validation and holdout test set from developmental dataset); and lack of full text. The proportion was determined by using 59 ML prediction models with only internal validation in orthopaedic surgical outcome published up until 18th June 2020 – previously identified by our group. Model performance was evaluated using discrimination, calibration, and decision-curve analysis. The TRIPOD guidelines assessed transparent reporting.

### Results

We included 18 studies externally validating 10 different ML prediction models of the 59 available ML models after screening 4682 studies. All external validations identified in this review retained good discrimination. Other key performance measures were only provided in 3 studies, rendering overall performance evaluation difficult. The overall median TRIPOD completeness was 61% (IQR, 43-89%), with 6 items being reported in less than 4/18 of the studies.

### Conclusion

Most current predictive ML models are not externally validated. The 18 available external validation studies were characterized by incomplete reporting of performance measures, limiting a transparent

examination of model performance. Further prospective studies are needed to validate or refute the myriad of predictive ML models in orthopedics while adhering to existing guidelines. This ensures clinicians full advantage of validated and clinically implementable ML decision tools.

# INTRODUCTION

Multiple machine learning (ML) algorithms have been recently developed for prediction of outcomes in orthopedic surgery. A recent systematic review demonstrated that 59 models are currently available covering a wide variety of surgical outcomes, such as survival, postoperative complications, hospitalization, or discharge disposition to aid clinical decision-making.[1] However, it is imperative that these models are accurate, reliable, and applicable to patients outside the developmental dataset. Even though internal validation studies regularly report good performance, these results are often too optimistic as performance on external validation worsens due to initial overfitting.[2,3]

External validation refers to assessing the model's performance on a dataset that was not used during development. Testing the developed model on independent datasets addresses the aforementioned concerns of internal validation, including: the generalizability of the model in different patient populations, shortcomings in statistical modelling (e.g., incorrect handling of missing data), and model overfitting.[2,4] Therefore, external validation is essential before a model can be used in routine clinical practice.

Although a growing number of ML prediction models are being developed in orthopedics, no overview exists of the number of available ML prediction models that are externally validated, how they perform in an independent dataset, and what the transparency of reporting is of these external validation studies. Therefore, we assessed the proportion, performance and transparent reporting of externally validated ML prediction models in orthopedic surgery, using the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines.

# METHODS

### Systematic Literature Search

Adhering to the 2009 PRISMA guidelines, this review was registered online at PROSPERO.[5] A systematic search was conducted in PubMed, Embase and Cochrane up until November 17th, 2020.

Three different domains of medical subject headings (MeSH) terms and keywords were combined with 'AND', and within domains the terms were combined with 'OR'. The 3 domains included words related to orthopedics, ML, and external validation. In addition, we searched the first and last authors from the 59 ML prediction models previously identified in a systematic review by our study group combined with the domain "machine learning" (Appendix 1). Two authors (NDK, PKT) independently screened all titles and abstracts. All references of the included studies were examined for relevant studies not identified by the initial search. The final list of included studies was sent to all coauthors, all of whom had worked with and/or published ML prediction models in orthopedics for a last check of potentially missed studies (Figure 1).

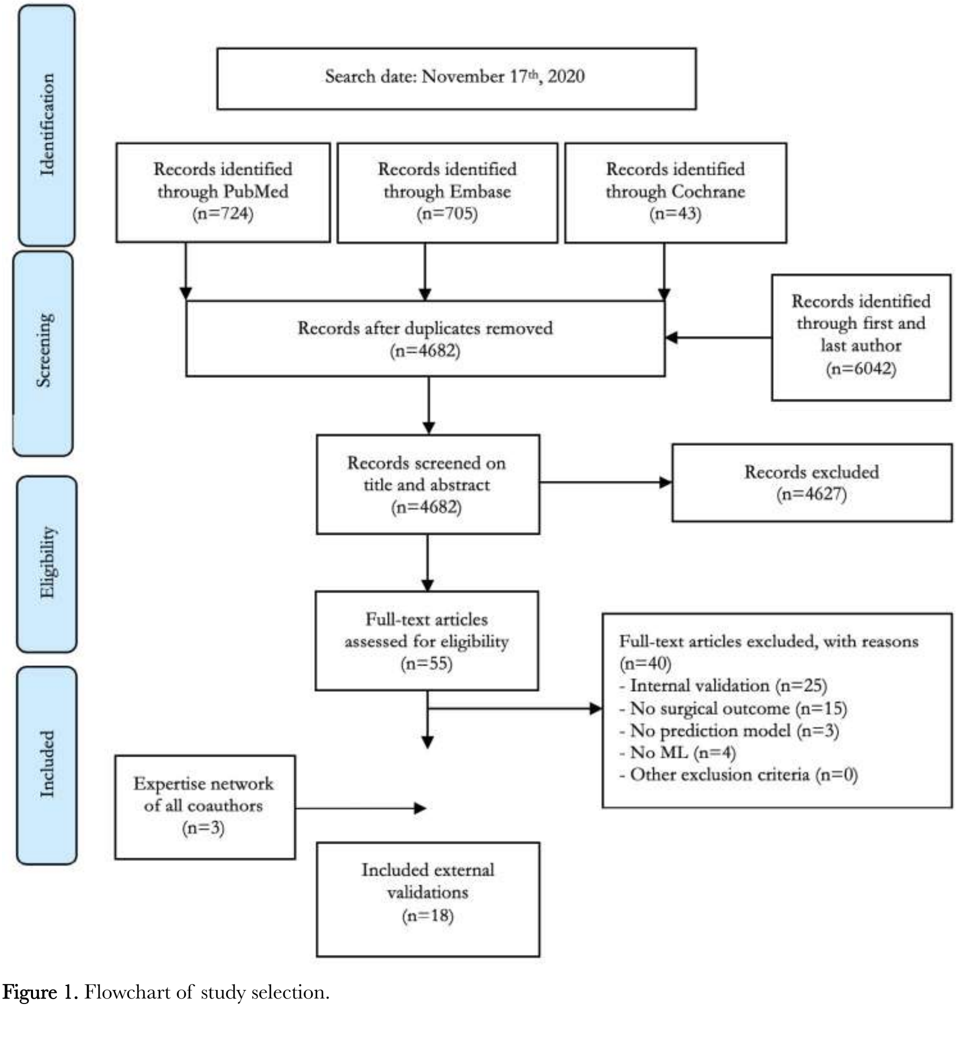

Identification
Screening
Eligibility
Included
Search date: November 17th, 2020
Records identified through PubMed (n=724)
Records identified through Embase (n=705)
Records identified through Cochrane (n=43)
Records identified through first and last author (n=6042)
Records after duplicates removed (n=4682)
Records screened on title and abstract (n=4682)
Records excluded (n=4627)
Full-text articles assessed for eligibility (n=55)
Full-text articles excluded, with reasons (n=40)
- Internal validation (n=25)
- No surgical outcome (n=15)
- No prediction model (n=3)
- No ML (n=4)
- Other exclusion criteria (n=0)
Expertise network of all coauthors (n=3)
Included external validations (n=18)

**Figure 1.** Flowchart of study selection.

### Eligibility Criteria

Inclusion criteria were: external validation; prediction models based on ML; and orthopedic surgical outcome (defined as any outcome after musculoskeletal surgery). Exclusion criteria were: non-ML prediction model (e.g., standard logistic regression); internal validation (e.g., cross-validation and holdout test set from developmental dataset); lack of full text; conference abstracts; animal studies; and languages other than English, Spanish, German, or Dutch. We considered advanced logistic regression methods as ML algorithms such as penalized LR (LASSO, ridge, or elastic-net), boosted LR and bagged LR.

### Data Extraction

Data extracted from each study were: year of publication; 1st author; disease; type of surgery; prospective study design; level of care from which the dataset originates (e.g. tertiary); country; type of ML algorithm (e.g. Bayesian Belief Network); sample size; input features; predicted outcome; time points of outcome; performance measures according to the ABCD approach[6] (A=calibration-in-the-large, or the model intercept; B=calibration slope; C=discrimination, with an area under the curve (AUC) using evaluation metrics of receiver operating characteristic (ROC) curves or precision-recall (PR) plots; D=decision-curve analysis); mention of guideline adherence; TRIPOD items[4]; and PROBAST domains.[7] Data were extracted from the largest cohort when multiple cohorts were present and the best performing model if a study reported results for multiple outcomes (e.g., 90-day and 1-year survival). Performance measures of the developmental study were extracted to compare with the results of external validation. 2 reviewers (OQG, BJJB) independently extracted all data and disagreements were discussed with a third reviewer present (PTO) until consensus was achieved.

### TRIPOD and PROBAST

The TRIPOD guidelines were simultaneously published in 11 leading medical journals in January 2015[4]. Although various other guidelines exist[8,9], we deemed the TRIPOD guidelines essential for transparent reporting requirements, which is imperative when judging the validity and applicability of a prediction model. Also, the TRIPOD guidelines were developed entirely for transparent reporting of prognosis or diagnosis prediction model studies (Appendix 2 and 3).

The PROBAST assesses the risk of bias of a study that validates a prognostic prediction model.[7] It is specifically designed to grade studies included in a systematic review, 4 domains are assessed for risk of bias: (1) participants; (2) predictors; (3) outcome; (4) and analysis (Appendix 4).

### Statistical Analysis

The proportion of externally validated ML prediction models in orthopedic surgical outcome was calculated by dividing 59 models by the externally validated models identified through this current study. Our group previously found 59 ML prediction models using only internal validation meeting the same criteria (except the criterium was “developmental” instead of “external validation”) in a systematic search dated up until June 18th, 2020 (Groot et al. 2021, Ogink et al. 2021). Of the identified external validation studies, we determined how many unique models were externally validated, as 1 model can be externally validated multiple times with different datasets. 1 incremental value study was found, which also reported on external validation. Only the external validation part was assessed.

Performance measures were extracted and expressed as they were originally reported.[6] No meta-analysis could be performed because of obvious heterogeneity between studies. Adherence to the TRIPOD guidelines and PROBAST domains were expressed in percentages and visualized by graphs.

We used Microsoft Excel Version 19.11 (Microsoft Inc, Redmond, WA, USA) to extract data using standardized forms, and to create all figures and tables, and Mendeley Desktop Version 1.19.4 (Mendeley Ltd, London, UK) as reference software.

# RESULTS

4,682 unique studies were identified of which 15 remained after full-text screening. 3 studies missed by the search were added by the coauthor's expertise network. None of the external validations used a prospective cohort and 12/18 investigated survival in bone oncology (Table 1). 6/18 mentioned adherence to the TRIPOD guidelines, but none included the actual checklist. All studies were affiliated with 6 institutions of which 7/18 with PATHFx and 5/18 with SORG. 17/18 had at least 1 author who was also an author on the paper that developed the model being evaluated. 9/18 of the studies reported on both development and external validation in the same paper; the other 9 only reported on external validation. All of the ML prediction models were freely available at www.pathfx.org, www.sorg-ai.com, safetka.net/, http://med.stanford.edu/s-spire/Resources/clinical-tools-.

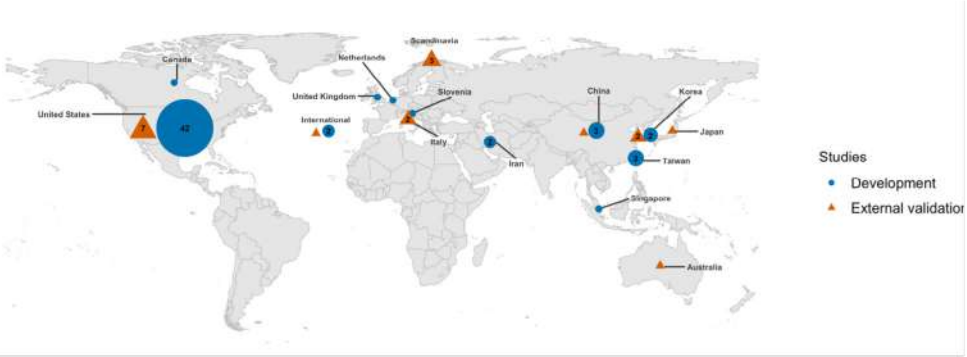


**Figure 2.** Distribution of development and external validation studies. All the developmental studies that were externally validated except two South-Korean were built on American datasets, unlike the origin of the external validation studies. Symbols without a number correspond with 1 study. Studies that included both development and external validation within the same study were counted twice in the figure according to where both datasets originated from.

html, and https://github.com/JaretK/NeuralNetArthroplasty. 17 datasets were used because 3 studies used 1 Scandinavian dataset and 1 study included 2 validation registry cohorts (Table 2). 14/17 of the datasets originated from hospitals, the other 3 were from a registry. The median sample size of the external validation datasets was 274 patients (IQR, 178-552) and 7/17 were American datasets (Figure 2).

**Table 1.** External validation characteristics of orthopaedic surgical outcome prediction studies (n=18)

| *Studies, year* | *Disease condition* | *Operation* | *ML model* | *Prospective database* | *Output* | *Input predictors* | *Number of patients* | *Adherence to a guideline* |
|---|---|---|---|---|---|---|---|---|
| Anderson, 2020 | Pathological fractures | nos | BBN | no | Survival | Clinical | 197 | TRIPOD |
| Bongers, 2019 | Extracranial chondrosarcoma | nos | BPM | no | Survival | Clinical | 179 | none |
| Bongers1, 2020 | Extracranial chondrosarcoma | nos | BPM | no | Survival | Clinical | 464 | TRIPOD |
| Bongers2, 2020 | Bone metastases (spine) | nos | SGB | no | Survival | Clinical | 200 | TRIPOD |
| Forsberg, 2012 | Bone metastases (extremities) | nos | BBN | no | Survival | Clinical | 815 | none |
| Forsberg, 2017 | Bone metastases | nos | BBN | no | Survival | Clinical | 815 | TRIPOD |
| Harris, 2019 | nos | Elective TJA | LASSO | no | Survival; complications | Clinical | 70569 | none |
| Huang, 2019 | Non-metastatic chondrosarcoma | nos | LASSO | no | Survival | Clinical, Surgical | 72 | none |
| Jo, 2019 | nos | TKA | GBM | no | Transfusion | Clinical, Surgical | 400 | none |
| Karhade, 2019 | Bone metastases (spine) | nos | SGB | no | Survival | Clinical | 176 | TRIPOD |
| Ko, 2020 | nos | TKA | GBM | no | Acute kidney injury | Clinical, Surgical | 455 | none |
| Meares, 2019 | Bone metastases (femoral) | nos | BBN | no | Survival | Clinical | 114 | none |
| Ogura, 2017 | Bone metastases | nos | BBN | no | Survival | Clinical | 261 | none |
| Overmann, 2020 | Bone metastases (extremities) | nos | BBN | no | Survival | Clinical | 815 | none |
| Piccioli, 2015 | Bone metastases | nos | BBN | no | Survival | Clinical | 287 | none |
| Ramkumar1, 2019 | Osteoarthritis | THA | ANN | no | LOS; discharge disposition | Clinical | 2771 | none |
| Ramkumar2, 2019 | Osteoarthritis | TKA | ANN | no | LOS; discharge disposition | Clinical | 4017 | none |
| Stopa, 2019 | Lumbar disc disorder | Decompression or fusion | NN | no | Nonhome discharge | Clinical, Surgical | 144 | TRIPOD |

*ML=machine learning; nos=not otherwise specified; TJA=total joint arthroplasty; TKA=total knee arthroplasty; THA=total hip arthroplasty; BBN=Bayesian Belief Network; NN=neural network; BPM=Bayes Point Machine; SGB=Stochastic Gradient Boosting; LASSO=least absolute shrinkage and selection operator; GBM=gradient boosting machine; LOS=length of stay; TRIPOD=Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis*

Table 2. Characteristics of hospital setting and years of enrollment from external validation

| Model or institution | Similar authors development and validation | *External validation* | | | | | |
|---|---|---|---|---|---|---|---|
| | | First author, year | Country | Tertiary | Hospitals | Registry | Years of enrollment |
| Cleveland | yes | Ramkumar1, 2019 | USA | mixed | 11 | no | 2016-2018 |
| Cleveland | yes | Ramkumar2, 2019 | USA | mixed | 11 | no | 2016-2018 |
| BETS/ PATHFx 1.0 | yes | Forsberg, 2012 | Scandinavia | yes | 8 | no | 1999-2009 |
| PATHFx 1.0 | yes | Piccioli, 2015 | Italy | yes | 13 | no | 2010-2013 |
| PATHFx 1.0 | yes | Forsberg, 2017 | Scandinavia | yes | 8 | no | 1999-2009 |
| PATHFx 1.0 | yes | Ogura, 2017 | Japan | yes | 5 | no | 2009-2015 |
| PATHFx 1.0 | no | Meares, 2019 | Australia | unknown | 1 | no | 2003-2014 |
| PATHFx 2.0 | yes | Overmann, 2020 | Scandinavia | yes | 8 | no | 1999-2009 |
| PATHFx 3.0 | yes | Anderson, 2020 | Multinational | yes | multiple | IBMR* | 2016-2018 |
| SafeTKA | yes | Jo, 2019 | unknown | unknown | 1 | no | unknown |
| SafeTKA | yes | Ko, 2019 | South-Korea | yes | 1 | no | 2018-2019 |
| SORG | yes | Bongers, 2019 | USA | yes | 2 | no | 1992-2013 |
| SORG | yes | Bongers1, 2020 | Italy | yes | 1 | no | 2000-2014 |
| SORG | yes | Karhade, 2019 | USA | yes | 1 | no | 2003-2016 |
| SORG | yes | Bongers2, 2020 | USA | yes | 1 | no | 2014-2016 |
| SORG | yes | Stopa, 2019 | USA | yes | 1 | no | 2013-2015 |
| Stanford | yes | Harris, 2019 | USA | mixed | multiple | VASQIP | 2005-2013 |
| Zhengzhou | yes | Huang, 2019 | China | yes | 1 | no | 2011-2016 |

*Nine studies include both the development and external validation (noted as the "same")*
*BETS=Bayesian Estimated Tools for Survival; SORG=Spinal Oncology Research Group; USA=United States of America; NSQIP=National Surgical Quality Improvement Program; SEER=Surveillance, Epidemiology, and End Results; IBMR=International Bone Metastasis Registry; NIS=National Inpatient Sample; VASQIP=Veterans Affairs Surgical Quality Improvement Program*

### Proportion

This systematic review identified 18 external validation studies of ML models predicting outcomes in orthopedic surgery. In these 18 external validation studies, 10 unique ML prediction models were validated as 2 models were validated twice, and 1 model 7 times as it was validated and updated multiple times with distinct datasets. Therefore, 10/59 of the ML models predicting outcomes in orthopedic surgery published up until June 18th, 2020 were externally validated. Of the 10 models, 3 were externally validated with patients from another country than the developmental cohort, including 1 model by 4 different countries.

and corresponding developmental studies.

| *Development* | | | | | |
|---|---|---|---|---|---|
| First author, year | Country | Tertiary | Hospitals | Registry | Year of enrollment |
| Same | USA | mixed | multiple | NIS | 2009-2011 |
| Same | USA | mixed | multiple | NIS | 2009-2013 |
| Forsberg, 2011 | USA | yes | 1 | no | 1999-2003 |
| Forsberg, 2011 | USA | yes | 1 | no | 1999-2003 |
| Same | USA | yes | 1 | no | 1999-2003 |
| Forsberg, 2011/2017 | USA | yes | 1 | no | 1999-2003 |
| Same | USA | yes | 1 | no | 1999-2003 |
| Same | USA | yes | 1 | no | 1999-2003, 2015-2018 |
| Same | South-Korea | yes | 1 | no | 2012-2018 |
| Same | South-Korea | yes | 2 | no | 2012-2019 |
| Thio, 2018 | USA | mixed | multiple | SEER | 2000-2010 |
| Karhade, 2019 | USA | yes | 2 | no | 2000-2016 |
| Karhade, 2018 | USA | mixed | multiple | NSQIP | 2011-2016 |
| Same | USA | mixed | multiple | NSQIP | 2013-2014 |
| Same | USA | mixed | multiple | SEER | 2005-2014 |

**This study also included an external validation on a second registry cohort of 192 patients from the Military Health System Data Repository*

### Performance

All studies reported the ROC AUC which retained good discriminative ability with a value greater than 0.70 and/or less than 0.10 decreased performance compared with the corresponding development study (Appendix 5 and 6). No PR AUC evaluation metrics were provided, despite 3/18 of the datasets consisting of imbalanced class distribution in which the ratio events:non-events were greater than 1:10. Calibration intercept and slope, or curve were reported in 7/18. 5/18 reported calibration slope or curves that showed overall underfitting of the data. Decision curve analyses were provided in 9/18, all of which illustrated that the prediction models were suitable for clinical use.

### TRIPOD and PROBAST

The overall median completeness of the TRIPOD items was 61% (IQR, 43-90%; Figure 3). All method items adhered to a median completeness of 56% (IQR, 44-72%) and all results items to a median of 42% (IQR, 22-61%). 6 items were reported in more than 16 studies including 3 discussion items (Table 3 and Appendix 7). 6 items were reported in less than 4 studies, including details of abstract, participant selection, and reporting key performance measures.

Participants selection (domain 1) was considered unclear risk of bias in 10 studies because no information was provided on the inclusion and exclusion of patients (Figure 4). Predictors (domain 2) were deemed low risk of bias in 16 studies, as 2 studies were unclear in their predictor's definitions and assessment. Outcome (domain 3) was rated high risk of bias in 2 studies as they determined survival not in a similar way for all patients by assigning "death" to all patients lost to follow-up. 2 additional studies in the outcome domain were rated unclear risk of bias because it was difficult to discern if they used the same postoperative complication definitions for both the development and external validation study. Analysis (domain 4) was rated high risk of bias in 17 studies, mainly due to small sample sizes with less than 100 events in the outcome group or no calibration metrics. The overall judgement of risk of bias for the 18 studies was high in 17 studies and low in 1 study, as only 1 study scored "low risk of bias" across all 4 domains.

**Table 3.** Sorted by completeness of above 90% reporting and under 25% of individual

| *Complete TRIPOD reporting >90%* | | |
|---|---|---|
| *Item* | *Description* | *% (n)* |
| 3a | Explain the medical context (including whether diagnostic or prognostic) and rationale for validating the multivariable prediction model, including references to existing models. | 100% (18) |
| 4a | Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the validation data set. | 100% (18) |
| 19b | Give an overall interpretation of the results considering objectives, limitations, results from similar studies and other relevant evidence. | 100% (18) |
| 22 | Give the source of funding and the role of the funders for the present study. | 100% (18) |
| 6b | Report any actions to blind assessment of the outcome to be predicted. | 94% (17) |
| 19a | Discuss the results with reference to performance in the development data, and any other validation data. | 94% (17) |

*TRIPOD=Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.*

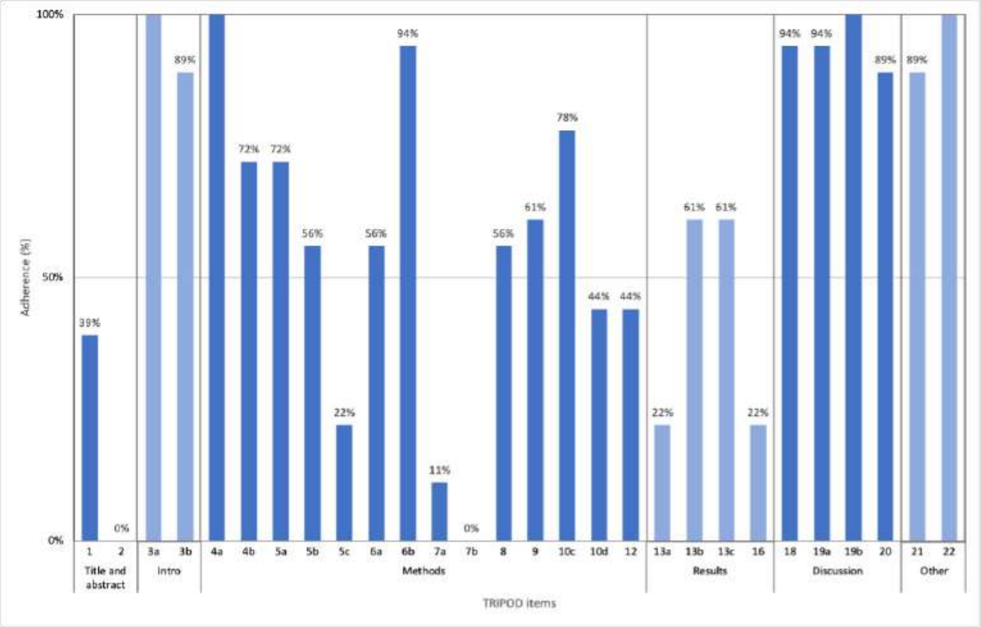


**Figure 3.** Overall adherence to each TRIPOD item (n=18).

TRIPOD items.

| *Complete TRIPOD reporting < 25%* | | |
|---|---|---|
| *Item* | *Description* | *% (n)* |
| 2 | Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions. | 0% (0) |
| 7b | Report any actions to blind assessment of predictors for the outcome and other predictors. | 0% (0) |
| 7a | Clearly define all predictors used in validating the multivariable prediction model, including how and when they were measured. | 11% (2) |
| 5c | Give details of treatments received, if relevant. | 22% (4) |
| 13a | Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful. | 22% (4) |
| 16 | Report performance measures (with confidence intervals) for the prediction model (results). | 22% (4) |

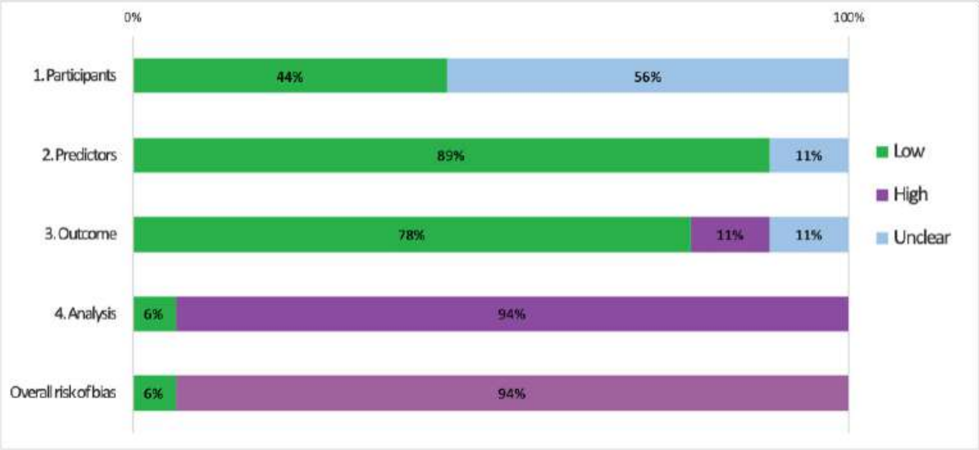


**Figure 4.** PROBAST results for all four domains and overall judgement (n=18).

# DISCUSSION

The focus on developing and publishing ML prediction models has led to an increasing body of studies. Yet, it is of equal importance to externally validate these models, as the TRIPOD states in their guidelines: *"external validation is an invaluable and crucial step in the introduction of a new prediction model before it should be considered for routine clinical practice."* Although the external validation studies identified in this review retained good discriminatory performance and overall adhered well to the TRIPOD guidelines, only 10/50 of the ML models predicting orthopedic surgical outcome published up until June 2020 have been externally validated. Skepticism of these non-externally validated models is necessary and an increased effort in externally validating existing models is required to realize the full potential of ML prediction models.

This study has several limitations. First, studies meeting the selection criteria may have been missed. However, we believe this was unlikely as we used 4 different search strategies. In addition, we believe that any missed studies would not have had a profound impact on the review's message as the percentage of externally validated models was well below 20%. Second, 5 of the 18 included studies originated from the authors' institution (SORG) and the reviewers may have been biased assessing them. To account for this potential bias, the second reviewer (BJJB) was not affiliated with the institution, the PI was not present during the consensus meetings, and an online PROSPERO protocol was registered. Third, publication bias may have occurred as successful external validations may be published more often. The performance results presented in this review may therefore be too optimistic and the number of studies externally validated too pessimistic. Studies demonstrating poorer performing models are part of the implementation process and ideally should be equally embraced by journals as high-performing models. In addition, the AUCs presented in 3 studies may have been too optimistic as they used ROC metrics on imbalanced datasets. Future studies should provide PR AUC metrics for datasets with an imbalanced class distribution [11]. Fourth, the presented low percentage of ML prediction models externally validated may have been unfair, as 20 ML models have been developed and published in the last year and external validation studies are time consuming. However, excluding the studies published in the last year to correct for this delay, still only yielded a disappointing 18/39 of ML prediction models that were externally validated. In addition, not all published ML-models are for deployment as we are still exploring the potentials of ML and therefore publications' primary motivation may be exploring the space of ML. Instead of externally validating these models, online tests should be provided where users can assess themselves how the ML models behave in different settings and parameters. Unfortunately, over half of the ML development studies did not provide online calculators, algorithms and/or open access (Ogink et al., 2021). Future ML studies should place more emphasis on providing easy to access means where outside users can assess model performance and behavior themselves. Fifth, various reporting guidelines exist such as STROBE and JMIR Guidelines for Developing and Reporting Machine Learning Models in Biomedical Research.[8,9] However, we used the TRIPOD guidelines to assess the transparent reporting as this guideline was explicitly developed to cover the development and validation of prediction models for prognosis.[4] To improve upon these guidelines, the TRIPOD authors are currently developing a TRIPOD-AI version specifically for reporting of AI prediction models.[12] Sixth, the guidelines are endorsed by 21 medical journals, of which only 1 is orthopedic (Journal of Orthopedic & Sports Physical Therapy). Since none of the studies were published in journals that officially endorsed the TRIPOD, it may be unfair to expect compliance to these guidelines. However, we believe that the TRIPOD guidelines present a high-quality benchmark for assessing transparent reporting, which is necessary for externally validating existing models and creating clinically implementable ML prediction models. Despite these limitations, our review provides valuable insights in the amount and transparent reporting of current ML external validations in orthopedics surgical outcome prediction.

A disappointingly low 10/59 of the current available ML prediction models were externally validated in orthopedic surgical outcome with none of the datasets being prospective. Prospectively testing the performance of ML models under real-world circumstances is an essential step towards integrating these models into the clinical setting and evaluating the impact on healthcare.[4] In addition, increased effort towards external validation on patient data from distinct geographic sites is needed as the generalizability of models to other countries may be affected by differences in healthcare systems, predictor measurements, and treatment strategies.[13] Although the recent surge of ML models in orthopedics is exciting, it is critical that these models are tested with external, real-world, operational data in different geographical settings before the orthopedic community can fully embrace the models in clinical practice.

The external validations identified in this review retained good discrimination. Other key characteristics recommended to evaluate a model's performance such as calibration, and decision-curve analysis were inadequately or not reported, as observed here and in similar reviews.[2,14–16] Calibration measures were only provided in 7 of the 18 studies, preventing a transparent examination of the model performance across the range of predicted probabilities.[6] Lastly, and arguably more important than the other metrics, is clinical usefulness evaluated by decision-curve analysis.[17] All 9 of the 18 studies that reported a decision-curve analysis indicated that the models were suitable for clinical use. Importantly, these curves do not estimate the likelihood of the outcome, but rather illustrate when the model should and should not be used in certain clinical situations over a range of thresholds. Overall, only 3 studies provided all 4 key measures to reliably evaluate the performance, despite a substantial body of methodological literature and published guidance emphasizing the importance of these performance measures.[4,6,8,9] Clinical researchers should use proposed frameworks such as Steyerberg's ABCD approach to systematically report the performance of a validated model to allow accurate evaluation.[6]

An additional interesting find is that 17 of the 18 studies were conducted by authors involved in the development of the model. Authors evaluating their own model might be overly optimistic, selectively report the results to their own advantage, and even defer publication if the performance is poor.[3] Although validating one's model is an essential first step, ideally this should be done by researchers not affiliated with the developmental study.

Although the external validations fared better in overall TRIPOD adherence than their corresponding developmental studies, they too had numerous incomplete items. The abstract, for which complete reporting required information on 12 elements, was incomplete in all studies. Some basic key details such as defining predictor definitions, outcome or treatment elements were poorly reported, despite not being specific to ML external validation studies. Specifying and reporting performance measures were poorly done in over half of the studies. Despite 6 TRIPOD items scoring less than 25% (5 were methods/results), 11 items scored over 75%, which included mainly introduction and discussion items. This difference in adherence across sections perhaps illustrates that the orthopedic community comprehends the rationale, promise and limitations of ML prediction models, but proper knowledge of methodological standards to describe and evaluate external validations studies is lacking. Standardized reporting and adherence to peer-reviewed guidelines such as the TRIPOD guidelines will aid in the execution and reporting of external validation studies – resulting in validated ML prediction models that are reliable, accurate, and that adds to surgical decision making.[4]

The PROBAST domains identified 2 major concerns in addition to the TRIPOD items. First, little attention was given to the flow of patient selection, as none of the studies included a flow diagram of included and excluded patients. Possibly, studies purposely did not include flow diagrams or selection criteria to maintain the generalizability of the model to patients outside of the selection criteria, but studies should explicitly state this. Second, the sample sizes were often too small, as only 5 of the 17 validation datasets had more than 100 events in each outcome group. Previous studies have shown that calibration results are less reliable with datasets less than 100 outcome events.[18] In most circumstances, it would have been difficult to reach this number as the disease conditions were primarily bone oncology related. To address the issue of inadequate number of outcomes, multi-institutional collaboration is needed to achieve effective sample sizes to allow reliable external validations.

# CONCLUSION

Despite the evident importance of evaluating the performance of prediction models on unseen datasets, it is rarely done as institutions are protective of sharing their data and journals prefer publishing development studies. In addition, algorithms that perform poorly on external validation may be subject to publication bias. The handful of available external validation studies overall adhered well to the TRIPOD guidelines, but certain items that are essential for transparent reporting were inadequately reported or not reported at all, namely details of the abstract, participant selection, and key performance measures. An increased effort to externally validate existing models on large, prospective, geographically distinct datasets is required to ensure accurate and reliable validated ML prediction models. It will be difficult to achieve these types of datasets without multi-institutional collaboration across different geographic regions. We encourage researchers and institutions, from both within and outside the orthopedic ML community, to collaborate.

# REFERENCES

1. Ogink PT, Groot OQ, Karhade AV, et al. Wide range of applications for machine learning prediction models in orthopaedic surgical outcome: a systematic review. *Acta Orthop 2021;92.*

2. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol.* 2014;14:40.

3. Siontis GCM, Tzoulaki I, Castaldi PJ, et al. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol.* 2015;68(1):25–34.

4. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* 2015;13:1.

5. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Rev Esp Nutr Humana y Diet.* 2016;20(2):148–160.

6. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35(29):1925–1931.

7. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med.* 2019;170(1):51–58.

8. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res.* 2016;18(12):e323.

9. von Elm E, Altman DG, Egger M, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ.* 2007;335(7624):806–808.

10. Groot OQ, Ogink PT, Lans A, et al. Poor reporting of methods and performance measures by machine learning studies in orthopaedic surgery: a systematic review. *J Orthop Res.* 2021;1:21.

11. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One.* 2015;10(3):e0118432.

12. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet (London, England).* 2019;393(10181):1577–1579.

13. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med.* 2013;10(2):e1001381.

14. Collins GS, Mallett S, Omar O, et al. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med.* 2011;9:103.

15. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med.* 2012;9(5):1–12.

16. Tangri N, Kitsios GD, Inker LA, et al. Risk prediction models for patients with chronic kidney disease: a systematic review. *Ann Intern Med.* 2013;158(8):596–603.

17. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak.* 2006;26(6):565–574.

18. Vergouwe Y, Steyerberg EW, Eijkemans MJC, et al. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol.* 2005;58(5):475–483.

# SUPPLEMENTAL MATERIAL TO CHAPTER 18

**Appendix 1.** Search syntaxes for the PubMed, Embase, and Cochrane Library on November 17th, 2020

**Appendix 2.** The TRIPOD guidelines state 22 items that are considered essential for informative reporting of prediction models. We used 18 out of the 22 main items, because item 11 "Risk groups" was not applicable to any of the external validation studies; item 14" Model development" and item 15 "Model specification" were not relevant to this review; and item 17 "Model updating" was only done in one study. Certain main items consisted of multiple subitems; main items 1, 2, 8, 9, 12, 16, 18, 20, 21, and 22 consisted of no subitems; main items 3, 4, 6, 7, and 19 consisted of two subitems (denoted by letters "a" and "b"; e.g., 3a and 3b), main items 5 and 13 consisted of three subitems (e.g., 5a, 5b, and 5c), and main item 10 consisted of five subitems. However, the subitems 10a "handling of predictors", 10b "model-building procedures", and 10e "model-updating" were also not rated as they were not relevant to this review. Main items and subitems are called under the same nomenclature "items" in the manuscript. In total, 28 items could be rated. Overall TRIPOD completeness was calculated per study and each separate item.

**Appendix 3.** Each item may consist of multiple elements. Each element was rated as "yes," "no," "referenced," or "not applicable." For an item to be considered incomplete, only one of the elements needed to be rated as "no." For an item to be considered complete, all of the elements needed to be rated as "yes," "referenced," or at least one of the previous two with the others "not applicable." For example, item 7a "defining predictors" consisted of four elements; (1) were all predictors reported; (2) were the predictors' definitions clearly presented; (3) how were the predictors measured; and (4) when were the predictors measured. Item 7a was considered incomplete if only one of the four elements were rated "no".

**Appendix 4.** Four domains are assessed for risk of bias: (1) participants; (2) predictors; (3) outcome; (4) and analysis. Each domain has several signaling questions to guide the rater towards a judgement. The four domains are rated as "low," "high," or "unclear" risk of bias. "Unclear" indicates that the reported information is insufficient – no reliable judgement on low or high risk of bias can be made with the information provided. To adapt the PROBAST specifically to our study purposes, we assigned a high risk of bias for the analysis domain (1) if the sample size was too small for the suggested minimum of 100 events in each outcome group18, or (2) when performance measures were not assessed according to Steyerberg's structured stepwise ABCD approach.[6] The number of 100 events in each group was deemed essential for the reliable evaluation of calibration plots. The validity of a prediction model was ideally assessed by four key metrics to evaluate the performance: calibration slope and intercept (or calibration curve), discrimination with an AUC, and clinical usefulness, with decision-curve analysis. The ratings of all four domains resulted in an overall judgement about risk of bias. Low overall risk of bias was assigned if each domain scored low. High overall risk of bias was assigned if at least one domain was judged to be high risk of bias. Unclear overall risk of bias was noted if at least one domain was judged unclear and all other domains low. The risk of bias for the four domains and overall judgement were reported – not the signaling questions.

**Appendix 5.** Performance measure of external validation studies compared to developmental studies according to the ABCD approach.

**Appendix 6.** Performance measure of external validation studies according to the ABCD. All provided AUC were ROC-AUC.

**Appendix 7.** Completeness of reporting of individual TRIPOD items.

Supplemental material can be consulted online per the website of the journal and/or publisher.

CHAPTER

19

# DOES ARTIFICIAL INTELLIGENCE OUTPERFORM NATURAL INTELLIGENCE IN INTERPRETING MUSCULOSKELETAL RADIOLOGICAL STUDIES? A SYSTEMATIC REVIEW

Olivier Q. Groot*, Michiel E.R. Bongers*, Paul T. Ogink, Joeky T. Senders, Aditya V. Karhade, Jos A.M. Bramer, Jorrit-Jan Verlaan, Joseph H. Schwab

**Joint first authorship*

*Clinical Orthopaedics and Related Research 2020;478:2751-2764*

# ABSTRACT

### Background

Machine learning (ML) is a subdomain of artificial intelligence that enables computers to abstract patterns from data without explicit programming. A myriad of impactful ML applications already exists in orthopaedics ranging from predicting infections after surgery to diagnostic imaging. However, no systematic reviews that we know of have compared, in particular, the performance of ML models with that of clinicians in musculoskeletal imaging to provide an up-to-date summary regarding the extent of applying ML to imaging diagnoses. By doing so, this review delves into where current ML developments stand in aiding orthopaedists in assessing musculoskeletal images.

### Objectives

This systematic review aimed (1) to compare performance of ML models versus clinicians in detecting, differentiating, or classifying orthopaedic abnormalities on imaging by (A) accuracy, sensitivity, and specificity, (B) input features (for example, plain radiographs, MRI scans, ultrasound), (C) clinician specialties, and (2) to compare the performance of clinician-aided versus unaided ML models.

### Design

Systematic review

### Methods

PubMed, Embase, and the Cochrane Library were searched for studies published up to October 1, 2019, using synonyms for machine learning and all potential orthopaedic specialties. We included all studies that compared ML models head-to-head against clinicians in the binary detection of abnormalities in musculoskeletal images. After screening 6531 studies, we ultimately included 12 studies. We conducted quality assessment using the Methodological Index for Non-randomized Studies (MINORS) checklist. All 12 studies were of comparable quality, and they all clearly included six of the eight critical appraisal items (study aim, input feature, ground truth, ML versus human comparison, performance metric, and ML model description). This justified summarizing the findings in a quantitative form by calculating the median absolute improvement of the ML models compared with clinicians for the following metrics of performance: accuracy, sensitivity, and specificity.

### Results

ML models provided, in aggregate, only very slight improvements in diagnostic accuracy and sensitivity compared with clinicians working alone and were on par in specificity (3% (interquartile range [IQR] -2.0% to 7.5%), 0.06% (IQR -0.03 to 0.14), and 0.00 (IQR -0.048 to 0.048), respectively). Inputs used by the ML models were plain radiographs (n=8), MRI scans (n=3), and ultrasound examinations (n=1). Overall, ML models outperformed clinicians more when interpreting plain radiographs than when interpreting MRIs (17 of 34 and 3 of 16 performance comparisons, respectively). Orthopaedists and radiologists performed similarly to ML models, while ML models mostly outperformed other clinicians (outperformance in 7 of 19, 7 of 23, and 6 of 10 performance comparisons, respectively). Two studies evaluated the performance of clinicians aided and unaided by ML models; both demonstrated considerable improvements in ML-aided clinician performance by reporting a 47% decrease of misinterpretation rate (95% confidence interval [CI] 37 to 54; $p < 0.001$) and a mean increase in specificity of 0.048 (95% CI 0.029 to 0.068; $p < 0.001$) in detecting abnormalities on musculoskeletal images.

### Conclusion

At present, ML models have comparable performance to clinicians in assessing musculoskeletal images. ML models may enhance the performance of clinicians as a technical supplement rather than as a replacement for clinical intelligence. Future ML-related studies should emphasize how ML models can complement clinicians, instead of determining the overall superiority of one versus the other. This can be accomplished by improving transparent reporting, diminishing bias, determining the feasibility of implantation in the clinical setting, and appropriately tempering conclusions.

# INTRODUCTION

Artificial intelligence is the capability of computers to display intelligent behavior, as opposed to humans, who demonstrate natural intelligence.[1,2] Machine learning (ML) is a subdomain of artificial intelligence that enables computers to abstract patterns from data without explicit programming.[3,4] ML applications are rapidly entering clinical practice in a variety of domains ranging from diagnostic to prognostic purposes.[5–7] The two most common types of ML used in medicine are supervised and unsupervised ML.[8,9] Supervised learning requires both input variables and labeled outcomes. In this form of ML, the algorithms learn to map the relationships between the input variables and outcomes.[8,10] Examples include processing the input of plain radiographs to detect the presence or absence of a fracture, often performed by convolutional neural networks (Figure 1). Unsupervised learning, unlike supervised learning, only requires input variables.[8] The algorithm seeks to find unknown patterns in the dataset to structure the data, without reference to a known outcome.

Several ML models and applications already exist in orthopaedics.[11–23] Despite the number of available studies, few systematic reviews or meta-analyses have examined the quality, limitations, and potential of ML models versus clinicians. Our group conducted a similar study in a wide range of neurosurgical applications which suggested that ML outperformed humans using multiple input features including radiographic and clinical parameters.[24] However, this review lacked scrutiny of the differences in input features and subspecialties and an in-depth discussion of the potential of ML models in musculoskeletal imaging. The potential benefit of the implementation of ML models to assess radiographs in orthopaedics is especially worthwhile, as misinterpretation is the primary reason for malpractice claims and may lead to grave clinical consequences such as malunion or joint collapse.[25] Furthermore, the systematic neurosurgical review performed in 2016 does not reflect the current ML environment since novel techniques, new forms of knowledge, and additional explanatory methods are being developed exponentially rather than linearly. Recent nonorthopaedic high-profile studies published between 2017 and now such as Esteva et al.5, Ting et al.[26], Lundberg et al.[27], Tomašev et al.[28], Liang et al.[29], Lee et al.[30], Hollon et al.[31], and Milea et al.[32], have transformed our understanding of the potential for ML to aid or replace clinicians. These studies have compared the algorithms to clinical experts and shown that these algorithms are able to diagnose or predict better than experts in a fraction of the time. Updated studies in this growing field of ML applications in medicine will help us understand if ML changes our expectations for the role of clinicians in the future. To our knowledge, no systematic reviews have compared the performance of the currently available ML models to the performance of clinicians in musculoskeletal imaging.

In this systematic review, we therefore aimed: (1) to compare performance of ML models versus clinicians on detecting, differentiating, or classifying orthopaedic abnormalities on imaging by (A) accuracy, sensitivity, and specificity, (B) input features (for example, plain radiographs, MRI scans, ultrasound), (C) clinician specialties, and (2) compare performance of clinicians aided versus unaided by ML models.

# METHODS

### Systematic Literature Search

We performed a systematic search in PubMed, Embase, and the Cochrane Library for studies published up to October 1, 2019. The search syntax was built with the guidance of a professional medical librarian using synonyms for "machine learning" and all potential orthopaedic specialties (Appendix 1). Two reviewers (OQG, MERB) independently screened all titles and abstracts for eligible articles based on predefined criteria (detailed below). Full-text articles were evaluated, and the references of the identified studies were examined for potentially relevant articles that were not identified by the initial search. Disagreements were solved by a discussion in which two other authors (PTO, JHS) were involved to assess article inclusion, quality assessment, and data extraction, until there was a consensus. We adhered to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines for this review.[33]

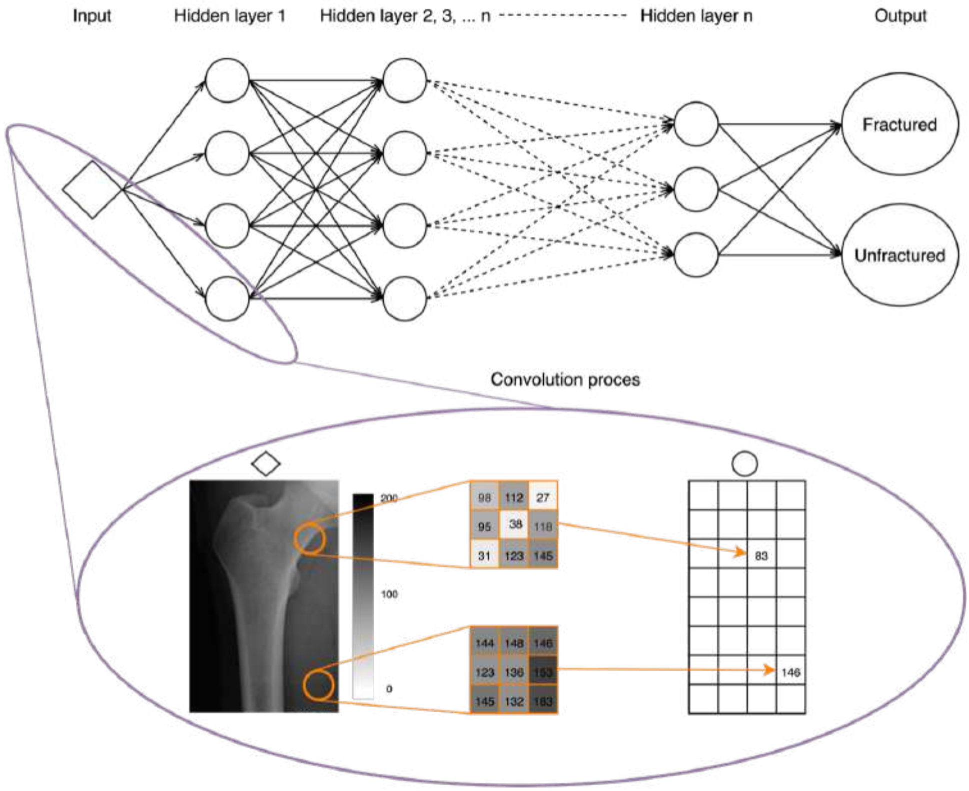


**Figure 1.** This figure shows a basic explanation of the most frequently used supervised learning algorithm—convolutional neural networks—for diagnosing orthopaedic conditions with imaging. A convolutional neural network transforms the input (for example, a plain radiograph of the femur) into one or more classification outputs (fracture or unfractured). The expanded box is a snapshot of the convolutional process, in which the input radiograph is processed into a matrix of pixel values (orange squares). After applying different filters developed in the training process, a single value is created in the output matrix (bottom right). This process is repeated in multiple hidden layers with different filters convolving across output matrices throughout hidden layers. Based on the connections and weights in the last hidden layer, the algorithm classifies the femur into fractured or not.

## Eligibility Criteria

Articles were included if they compared ML models head-to-head with clinicians in applications relevant to the orthopaedic patient population. We defined the orthopaedic patient population as patients with disorders of the bones, joints, ligaments, tendons, and muscles. All application domains such as diagnosis, prognosis, treatment, and outcome were included. In ML, the "ground truth" refers to the reference standard on which the model is trained and tested. This ground truth varied by article depending on its specific domain, including surgical or histologic confirmation in a radiologic classification task or the consensus of a panel of experts. We excluded studies that did not compare ML models and human performance, nonorthopaedic specialty studies, non-English-language studies, studies with no full text available, and nonrelevant article types, such as case reports, animal studies, and letters to the editor.

### Assessment of Methodological Quality

Two reviewers (OQG, MERB) independently appraised the quality of the included studies using predefined extraction sheets, based on the Methodological Index for Non-randomized Studies (MINORS) criteria.[34] We modified the seven-item MINORS checklist to make it applicable to our systematic review by including disclosure, study aim, input feature, ground truth, comparison between ML model and clinician, dataset distribution, performance metric, and description of the ML model. These eight items were scored on a 2-point scale: 0 (not reported or unclear) or 1 (reported and adequate).

After screening 6531 titles and abstracts, we assessed 40 full-text studies for eligibility, and ultimately 14 studies were included for critical appraisal (Figure 2). The study aim, inclusion and exclusion criteria for the input features, ML model used, and the human comparison group were clearly explained in all studies. The distribution of the dataset was clearly described in 11 studies; in the remainder of the studies, the dataset distribution was unclear or a test set was not used.[35–37] Disclosure was reported in 12 studies; thus, for two studies, conflicts of interest could not be evaluated.[38,39] The ground truth was not clearly described and clear performance metrics were missing in two studies.[37,39] This deviated considerably from existing reporting standards as it introduced bias by inadequate ground truth labeling and not providing transparent head-to-head comparison.[40] Thus, we excluded these two studies from this review. In total, 12 studies were included for quantitative synthesis (Appendix 2) and assessed using the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines for completeness of reporting of the ML model (Appendix 3). The TRIPOD guideline, which is a checklist of 22 items introduced in 2015, should be followed when reporting algorithm results.[40] This guideline is deemed essential for transparent reporting of study outcomes and guide developers of algorithms towards a more uniform reporting of their algorithm's performance.

### Data Extraction

Data obtained from each study were year of publication, output classes, performance measures, P-value of the difference in performance, input features, outcome measures, performance of ML, performance of the clinician, ML model, level of education of the human performer and (sub) specialization of the clinician, ground truth, size of the dataset, size of training set, validation method or size of the validation set, and size of the test set. For studies comparing multiple outcome measures between artificial and natural intelligence or comparing different groups of clinicians with ML models, each separate comparison was extracted.

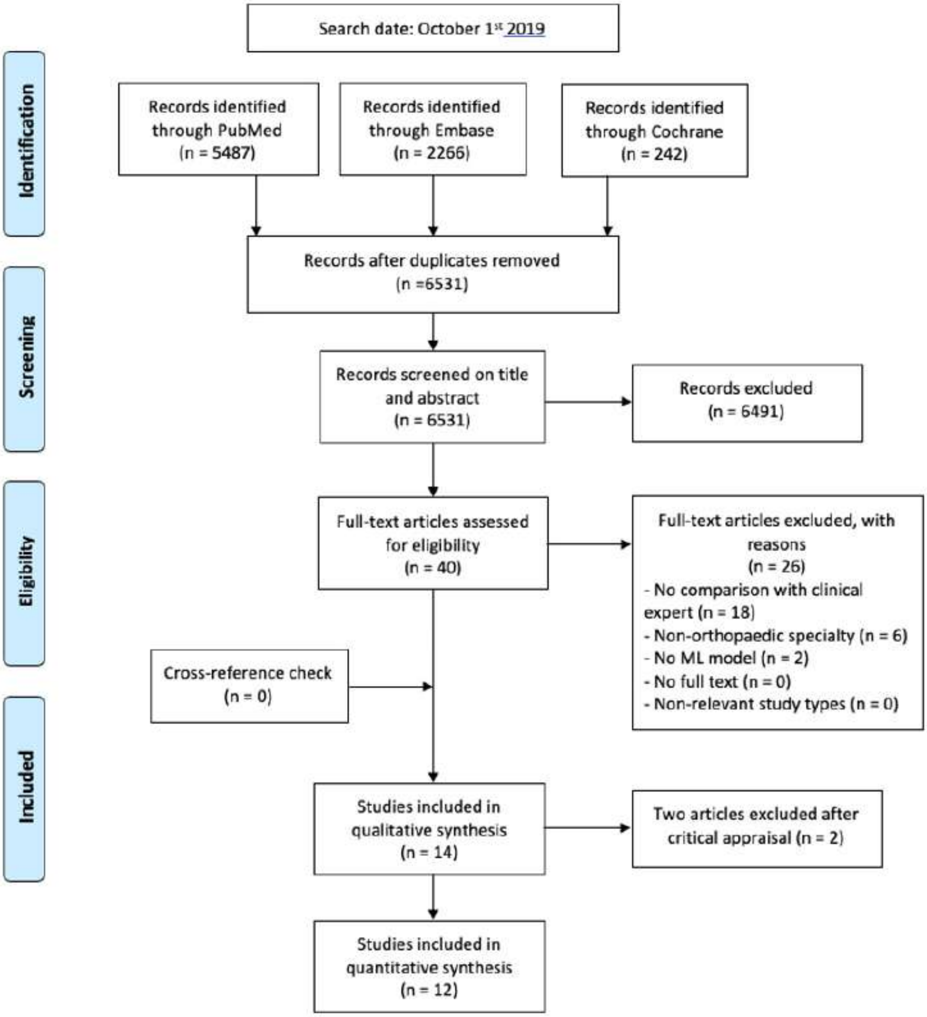

Search date: October 1st 2019
Identification
Records identified through PubMed (n = 5487)
Records identified through Embase (n = 2266)
Records identified through Cochrane (n = 242)
Screening
Records after duplicates removed (n =6531)
Records screened on title and abstract (n = 6531)
Records excluded (n = 6491)
Eligibility
Full-text articles assessed for eligibility (n = 40)
Full-text articles excluded, with reasons (n = 26)
- No comparison with clinical expert (n = 18)
- Non-orthopaedic specialty (n = 6)
- No ML model (n = 2)
- No full text (n = 0)
- Non-relevant study types (n = 0)
Cross-reference check (n = 0)
Included
Studies included in qualitative synthesis (n = 14)
Two articles excluded after critical appraisal (n = 2)
Studies included in quantitative synthesis (n = 12)

**Figure 2.** This Preferred Reporting Items for Systematic Reviews and Meta-analyses 2009 flow diagram shows how studies were systematically identified, screened, and included. After screening 6531 studies, 14 studies were critically appraised and ultimately 12 studies were included for quantitative synthesis.

## Study Characteristics

The median size of the training set was 1702 datapoints (interquartile range 337 to 16,075), that of the validation set was 334 datapoints (134 to 37,481), and that of the test set was 334 datapoints (155 to 2410). Five studies used cross-validation only instead of a separate validation set.[36,41–44] Two studies did not use a test set.[36,42] All studies used a binary assessment. No studies provided additional information (for example, physical examination findings) to either ML models or clinicians. No studies were designed as a prospective, randomized, controlled trial. None of the studies adhered to all TRIPOD checklist items.

Output classes for the 12 studies comparing ML models and humans were binary detection of fractures or other radiologic abnormalities (n=11)[35,36,38,42–49] or both detection and classification of the diagnosis (n=1).[41] Input features used by the ML models were plain radiographs (n=8)[38,41,44–49], MRI (n=3)[35,42,43], and ultrasound examinations (n=1).[36] A P-value was provided for 91% (52 of 57) of the outcome measures. Outcome measures accompanied by a P-value were used to assess the performance of ML models and clinicians; sensitivity and specificity (both 33% [17 of 52]), accuracy (31% [16 of 52]), and area under the receiver operating characteristic curve [AUC] (3.8% [2 of 52]). All ML models were supervised learning algorithms with the following two subtypes: convolutional neural networks (n=11)[35,38,41–49] and random forest ML (n=1).[36] All studies used publicly available pretrained models or data augmentation methods during training. Ground truth differed by study and was established by expert agreement with the aid of a more advanced radiographic modality (n=3)[38,41,47], expert agreement without the aid of a more advanced radiographic modality (n=6)[35,43-45,48,49], surgical or histologic confirmation (n=2)[42,46], and clinical diagnosis (n=1).[36]

The studies were also analyzed by the type of input feature used and by the specialty of the clinician expert. Input features could be divided into two main categories: plain radiographs and MRIs. The interpretation of plain radiographs by ML models was compared with that of clinicians in eight studies: detection of osseous abnormalities (n=8)[38,41,44–49] and fracture classification (n=1).[41] Detection of osseous abnormalities was the focus of seven studies, namely distal radius fractures[47,48], femoral neck fractures[46], intertrochanteric hip fractures[45], hip osteoarthritis[44], femoral head osteonecrosis[38], or any fracture in the hand, wrist, or ankle.[49] The detection and classification of proximal humerus fractures were investigated by one study.[41] MRI interpretation by ML models was compared with that of clinicians in three studies. The first study evaluated the detection of general abnormalities in the knee, ACL tears, and meniscal tears[35]; the second study focused on the ability to differentiate between tuberculous and pyogenic spondylitis[42]; and the third study evaluated the detection of cartilage lesions of the knee.[43] Ultrasound examination as an input feature was used in one study to distinguish between lateral epicondylosis and asymptomatic elbows.[36]

Assessing physicians were divided into three groups by their comparison with ML models: radiologists (6 of 12 studies)[35,36,38,42,43,47], orthopaedic surgeons (4 of 12)[41,45,47,49], and all others (5 of 12), including physiotherapists[36], general physicians[41,44], emergency medicine clinicians (consisting of physicians assistants and medical doctors)[48], and undergraduate students with different levels of education.[46]

### Statistical Analysis

Given the heterogeneity of the orthopaedic applications, no quantitative meta-analysis was performed. Because all 12 studies were of comparable quality and they all clearly included six of the eight critical appraisal items (study aim, input feature, ground truth, ML versus human comparison, performance metric, and ML model description), a quantitative summarization was provided by calculating the median absolute improvement. The median absolute improvement was determined by calculating the differences in performance metrics between the ML model and clinician for the most commonly used statistical measures of performance: accuracy, sensitivity, specificity, and AUC. The absolute median represents an overview of performance where positive and negative values correspond with superior performance of the ML model and clinician, respectively. No significance of any sort can be attributed to this summary metric. Accuracy refers to the proportion of total correct predictions among the total number of predictions, sensitivity refers to the proportion of true positive cases among the total number of positive cases, and specificity refers to the proportion of true negative cases among the total number of negative cases. AUC refers to the ability of the algorithm to discriminate between two classes ranging from 0 to 1.

Superior or inferior performance of the ML model versus that of clinicians was defined as a significant better or worse performance, respectively, according to the statistical tests used in the studies ($p < 0.05$). Equal performance was defined as a nonsignificant performance difference ($p > 0.05$). The sizes of the training, validation, and test sets are reported as percentages of the total dataset. We used Microsoft Excel Version 19.11 (Microsoft Inc, Redmond, WA, USA) and Stata® 14.0 (StataCorp LP, College Station, TX, USA) for the statistical analyses, and Mendeley Desktop Version 1.19.4 (Mendeley Ltd., London, UK) as reference management software.

# RESULTS

### Accuracy, Sensitivity, and Specificity

Machine learning models slightly outperformed clinicians working alone in detecting, differentiating, or classifying orthopaedic abnormalities on musculoskeletal imaging in diagnostic accuracy and sensitivity, and were on par in specificity. The median (range) absolute improvement values were 3.0% (-12% to 19%; IQR -2.0% to 7.5%)[35,41–47,49] for accuracy, 0.06 (-0.15 to 0.41; IQR -0.03 to 0.14) for sensitivity, and 0.00 (-0.15 to 0.13; IQR -0.048 to 0.048)[35,36,38,41–45,47,48] for specificity. The wide IQRs and ranges in all three performance measures narrow toward zero, which indicates that there was no strong difference between the performance of ML models and clinicians. The median absolute improvement in the AUC was not calculated because only four comparisons were provided.[36,38,42,45] The ML models performed better than clinicians in 38% of all performance measures (accuracy, sensitivity, and specificity; 20 of 52) and worse than clinicians in 3.8% (2 of 52); no difference was found in 58% (30 of 52) (Table 1).

### Results Stratified by Input Features

Machine learning models outperformed clinicians more frequently when interpreting plain radiographs than when interpreting MRIs. Interpretation of plain radiographs by ML models was better than that by clinicians in 17 of 34 of all performance measures (accuracy, sensitivity, and specificity) and worse in zero of 34; no difference was found in 17 of 34. On plain radiographs, ML models performed better than clinicians did in terms of detecting osseous abnormalities or classifying fractures in 13 of 28 all performance measures and 4 of 9, respectively; worse in 0 of 28 and 0 of 9, respectively; and no difference was found in 15 of 28 and 5 of 9, respectively. ML models were able to interpret MRIs better than clinicians in 3 of 16 of all performance measures and worse in 2 of 16; no difference was found in 11 of 16. Only one study evaluated ultrasound interpretations[36], and it showed no difference between ML models and clinicians in distinguishing between lateral epicondylosis and asymptomatic elbows.

### Results Stratified by Clinician Expert Specialty

Machine learning models performed similarly to radiologists and orthopaedists but better than all other clinicians. ML models performed better than clinicians in two specialist groups, orthopaedics and radiology, in 7 of 19 and 7 of 23 of all performance measures, respectively, and worse in 0 of 19 and 2 of 23, respectively; no difference was found in 12 of 19 and 14 of 23, respectively. ML models performed better than all other clinicians (physiotherapists, general physicians, emergency medicine clinicians, and undergraduate students) in 6 of all 10 outcome measures and worse in 0 of 10; no difference was found in 4 of 10.

### Results of Studies of ML Aiding Clinicians

Two studies evaluated the performance of clinicians aided and unaided by ML models; both demonstrated that clinicians aided by ML models outperformed clinicians unaided by ML. Lindsey et al.[48] showed that clinicians aided by ML models had improved performance in detecting wrist fractures compared with their non-aided performance. On average, clinicians had a relative proportional reduction of misinterpretation when aided by ML models of 47% (95% confidence interval 37 to 54; $p < 0.001$), compared to their non-aided performance. Bien et al.[35] evaluated the ML-aided and ML-unaided performance of clinicians in detecting general abnormalities and specific diagnoses on MRIs of the knee and found a mean increase in specificity of 0.048 for the aided detection of ACL tears (95% CI 0.029 to 0.068; $p < 0.001$).

**Table 1.** Performance of ML models and clinical experts

| Author[a] | Output | Input features | Outcome measures | ML models vs clinicians (95% CI)[#] |
|---|---|---|---|---|
| Adams | Detection of femur neck fracture | Radiography | Accuracy | 91% (86 to 95) vs 91% |
| Bien1[d] | Detection of general abnormality | MRI | Accuracy<br>Sensitivity<br>Specificity | 85% (78 to 90) vs 89% (87 to 91)<br>88% (80 to 93) vs 91% (88 to 92)<br>71% (50 to 86) vs 84% (78 to 89) |
| Bien2[d] | Detection of ACL tear | MRI | Accuracy<br>Sensitivity<br>Specificity | 87% (79 to 92) vs 92% (90 to 94)<br>76% (64 to 85) vs 91% (87 to 93)<br>97% (89 to 99) vs 93% (91 to 95) |
| Bien3[d] | Detection of meniscal tears | MRI | Accuracy<br>Sensitivity<br>Specificity | 73% (64 to 80) vs 85% (82 to 87)<br>71% (59 to 81) vs 82% (78 to 85)<br>74% (62 to 84) vs 88% (85 to 91) |
| Bureau | Differentiation of lateral epicondylosis and asymptomatic elbows | Ultrasound | AUC<br>Sensitivity<br>Specificity | 0.82 (0.80 to 0.85) vs 0.80 (0.66 to 0.94)<br>73% vs 68%<br>79% vs 86% |
| Chee | Detection of femoral head osteonecrosis | Radiography | AUC<br>Sensitivity<br>Specificity | 0.93 vs 0.91<br>79% vs 79%<br>95% vs 88% |
| Chung1 | Detection of proximal humerus fracture | Radiography | Accuracy<br>Sensitivity<br>Specificity | 96% (94 to 97) vs 85% (80 to 90)<br>99% (99 to 100) vs 82% (78 to 87)<br>97% (97 to 98) vs 94% (93 to 96) |
| Chung2 | Detection of proximal humerus fracture | Radiography | Accuracy<br>Sensitivity<br>Specificity | 96% (94 to 97) vs 93% (89 to 97)<br>99% (99 to 100) vs 93% (89 to 97)<br>97% (97 to 98) vs 97% (96 to 98) |
| Chung3 | Detection of proximal humerus fracture | Radiography | Accuracy<br>Sensitivity<br>Specificity | 96% (94 to 97) vs 93% (87 to 99)<br>99% (99 to 100) vs 96% (95 to 98)<br>97% (97 to 98) vs 98% (96 to 100) |
| Chung4 | Classifying normal, # of greater tuberosity, neck, 3-part, or 4-part | Radiography | Accuracy<br>Sensitivity<br>Specificity | 65% to 86% vs 32% to 82%<br>88% to 97% vs 33% to 69%<br>83% to 94% vs 84% to -94% |
| Chung5 | Classifying normal, # of the greater tuberosity, neck, 3-part, or 4-part | Radiography | Accuracy<br>Sensitivity<br>Specificity | 65% to 86% vs 43 to 90<br>88% to 97% vs 44% to % to 80%<br>83% to 94% vs 80% to 97% |
| Chung6 | Classifying normal, # of the greater tuberosity, neck, 3-part, or 4-part | Radiography | Accuracy<br>Sensitivity<br>Specificity | 65% to 86% vs 65% to 93%<br>88% to 97% vs 52% to 88%<br>83% to 94% vs 87% to 98% |
| Gan1 | Detection of distal radius fracture | Radiography | Accuracy<br>Sensitivity<br>Specificity | 93% (90 to 96) vs 94% (91 to 96)<br>90% (85 to 95) vs 93% (89 to 97)<br>96% (93 to 99) vs 95% (91 to 98) |
| Gan2 | Detection of distal radius fracture | Radiography | Accuracy<br>Sensitivity<br>Specificity | 93% (90 to 96) vs 84% (80 to 88)<br>90% (85 to 95) vs 81% (75 to 87)<br>96% (93 to 99) vs 87% (81 to 92) |

*Continued on next page*

| P-value | ML models vs clinicians | Total dataset | Training size[b] | Validation size[b]/method | Testing size[b] | Ground truth[c] |
|---|---|---|---|---|---|---|
| 0.999 | CNN vs<br>BSc students | 800 | 64% | 16% | 20% | Surgically confirmed |
| 0.301<br>0.620<br>0.344 | CNN vs<br>Rad | 1,370 | 91% | 9% | NA | Consensus of 3 Rad |
| 0.173<br>0.019<br>0.566 | CNN vs<br>7 Rad | 1,370 | 91% | 9% | NA | Consensus of 3 Rad |
| 0.082<br>0.619<br>0.019 | CNN vs<br>7 Rad | 1,370 | 91% | 0% | NA | Consensus of 3 Rad |
| NA<br>0.157<br>0.157 | RF vs<br>1 MSK Rad and 1 Phys | 54 | 100% | LOOCV | NA | Clinical diagnosis |
| NA<br>0.999<br>0.046 | CNN vs<br>2 Rad | 1,892 | 71% | 8% | 21% | Consensus of 2 Rad and MRI |
| <0.05<br><0.05<br><0.05 | CNN vs<br>28 GP | 1,891 | 90% | 10-FCV | 10% | Consensus of 2 Ortho, 1 Rad; CT for failed consensus |
| >0.05<br>>0.05<br>>0.05 | CNN vs<br>11 general Ortho | 1,891 | 90% | 10-FCV | 10% | Consensus of 2 Ortho, 1 Rad; CT for failed consensus |
| >0.05<br>>0.05<br>>0.05 | CNN vs<br>19 shoulder Ortho | 1,891 | 90% | 10-FCV | 10% | Consensus of 2 Ortho, 1 Rad; CT for failed consensus |
| 0.01<br><0.001<br>0.999 | CNN vs<br>28 GP | 1,891 | 90% | 10-FCV | 10% | Consensus of 2 Ortho, 1 Rad; CT for failed consensus |
| 0.094<br>0.001<br>0.999 | CNN vs<br>11 GP | 1,891 | 90% | 10-FCV | 10% | Consensus of 2 Ortho, 1 Rad; CT for failed consensus |
| 0.579<br><0.001<br>0.157 | CNN vs<br>19 shoulder Ortho | 1,891 | 90% | 10-FCV | 10% | Consensus of 2 Ortho, 1 Rad; CT for failed consensus |
| >0.05<br>>0.05<br>>0.05 | CNN vs<br>3 Ortho | 2,340 | 87% | 13% | 13% | Consensus of 3 Ortho and CT |
| <0.05<br><0.05<br><0.05 | CNN vs<br>3 Rad | 2,340 | 87% | 13% | 13% | Consensus of 3 Ortho and CT |

| | | | | |
|---|---|---|---|---|
| Kim | Differentiate tuberculous and pyogenic spondylitis | MRI | AUC<br>Accuracy<br>Sensitivity<br>Specificity | 0.80 (0.73 to 0.87) vs 0.73 (0.66 to 0.80)<br>76% (69 to 83) vs 70%<br>85% (75 to 92) vs 72%<br>68% (57 to 78) vs 69% |
| Lindsey[d] | Detection of wrist fracture | Radiography | Sensitivity<br>Specificity | 94% vs 81% (77 to 84)<br>95% vs 88% (85 to 90) |
| Liu | Detection of cartilage lesions within the knee joint | MRI | Accuracy<br>Sensitivity<br>Specificity | 84% vs 84%<br>82% vs 73%<br>87% vs 95% |
| Olczak | Detection of fracture: hand, wrist, ankle | Radiography | Accuracy | 83% (79 to 87) vs 82% (78 to 86) |
| Urakawa | Detection of intertrochanteric hip fracture | Radiography | AUC<br>Accuracy<br>Sensitivity<br>Specificity | 0.98 (0.97 to 0.99) vs 0.97 (0.95 to 0.98)<br>96% (93 to 98) vs 92% 89 to 95)<br>94% (90 to 97) vs 88% (83 to 93)<br>97% (95 to 99) vs 97% (95 to 98) |
| Xue | Detection of hip osteoarthritis | Radiography | Accuracy<br>Sensitivity<br>Specificity | 93% vs 88%<br>95% vs 100%<br>91% vs 78% |

Bold *values indicate that the difference between the performance machine learning models and clinicians was statistically significant ($p<0.05$). ML MSK=musculoskeletal; LOOCV=leave-one out cross validation; FCV=fold cross-validation; ED=emergency department; PA=physician assistant; for Bien et al., Chung et al., and Gan et al., for comparing multiple outcome measures between machine learning models and clinicians or compar (reference standard for machine learning models) varied between each study.* [d]*This study also used the measured performance of clinicians aided c*

# DISCUSSION

The availability of ML applications in the orthopaedic arena is increasing rapidly, but few studies have compared the performance of these models against their human counterparts. In 2017, we compared ML models and clinicians in the neurosurgical field and found that ML generally outperformed clinicians. However, that study was performed using not only imaging but also clinical input features in a wide variety of different ML models and was performed more than 3 years ago. Many advancements and novel techniques have transformed our understanding of the potential for ML since that time. Frequent determination of the advancements of ML in medicine and its performance compared with clinicians is important in this rapidly growing field. In fact, none of the included studies in this review had been published before our 2017 neurosurgical review. We found that ML models again outperformed clinicians more than clinicians outperformed ML models, but in aggregate these improvements were small. Also, clinicians aided by ML models performed better and faster compared to their non-aided performance. Machine learning models demonstrate great potential to improve the assessment of musculoskeletal imaging. However, significant hurdles–such as the lack of transparent reporting, inaccurate ground truth labeling, and transportability issues to the clinical non-research setting–must be overcome before clinicians can embrace ML models in daily practice.

This review has several limitations. First, summarizing the results with medians does not provide

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.281<br>0.002<br>0.002<br>0.317 | CNN vs<br>3 MSK Rad | 161 | 100% | 4-FCV | NA | Bacteriologic and/or histologic confirmation |
| NA<br>NA | CNN vs<br>39 ED (15 PAs; 24 MDs) | 135,845 | 80% | 17% | 3% | Subspecialized Ortho |
| 0.661<br><0.001<br><0.001 | CNN vs<br>Rad residents (2), fellows (2), staff (1) | 17,395 | 92% | 5-FCV | 8% | MSK Rad |
| NA | CNN vs<br>2 senior Ortho | 256,458 | 70% | 20% | 10% | Radiology report and three orthopaedists |
| <0.001<br><0.001<br><0.001<br>0.001 | CNN vs<br>5 Ortho | 3,346 | 80% | 10% | 10% | Ortho |
| 0.317<br>0.157<br>0.025 | CNN vs<br>3 Phys | 420 | 80% | 5-FCV | 20% | Consensus of 2 chief physicians |

*=machine learning; CNN=convolutional neural network; BSc=Bachelor of Science; NA=not available; RF=random forest; MD=medical doctors; Ortho=orthopedist; Rad=radiologist; GP=general practitioner; Phys=physiatrist. [a]Separate comparison were extracted ing different groups of clinicians with machine learning models. [b]Percentage of the total amount of the dataset. [c]The definition of ground truth nd unaided by machine learning models.*

adequate weight to each study based on quality and size. The size of studies ranged from 54 to 256,458 datapoints and no correction could be made for this imbalance. Two studies did not use a proper holdout test set[35,42], which could overestimate model performance as the data was used for both training and testing. Three studies assessed multiple outcome measures[35,41,47], resulting in an overrepresentation of these performance measures. Ideally, randomized controlled trials ensure fair comparison between ML models and clinicians, but to date, only two of these randomized trials exist.[50,51] However, to justify the data pooling, all included studies were of comparable high quality—maximum score on six of the eight critical appraisal items—and randomized clinical trials in ML models are not (yet) widely accepted. Second, our group conducted a similar review in 2017 in the field of neurosurgery[24]; there were no overlapping studies between both reviews. Third, ground-truth establishment differed throughout the studies, ranging from surgical or histologic confirmation to expert consensus. Some models could therefore have been trained on datasets containing human errors, leading to an overestimation of the clinician's performance. For example, an incorrectly labeled ground truth can lead to incorrect training of the algorithm, thereby falsely decreasing the algorithm's performance. If the clinician also does not assume that a fracture is present, his or her performance will falsely increase. In this review, all studies used relatively accurate ground truth labels such as data labeled by experts or histopathological confirmation compared with more error-prone radiology reports that may have been dictated by inexperienced junior residents. Therefore, the underestimation of the performance metrics of the ML models are of limited proportion.

Fourth, positive publication bias may have occurred because studies that reported the favorability of ML models may have been published more frequently. Additionally, all reviewed studies included comparisons in imaging, specifically in settings where ML models currently show the most promising results in multiple disciplines and are expected to outperform clinicians.[6,24] The superiority of ML models might therefore be overestimated and only applicable to imaging tasks, especially because it constitutes only one of the clinician's many specific tasks. It is reasonable to expect many trials in the near future to provide a more accurate comparison between ML models and clinicians as algorithm validation, implementation, and overall acceptance is increasing in clinical care. Fifth, the performance of the ML models could have been overestimated in studies that did not use a proper independent test set. Further, studies differed in the amount of analyses and outcome measures, which could have caused overrepresentation of some studies. No uniform comparison could have been made to prevent this overrepresentation because there was heterogeneous reporting of outcome measures. Furthermore, a P-value was not provided for four of 57 outcome measures. All four showed that the ML models had superior performance, and in these cases, the strength of the ML models might have been underestimated.[36,38,48] Sixth, the AUC was provided in only four studies with two P-values, making a comparison unwarranted. However, binary predictions were made in all studies, making this limitation less problematic. Seventh, because all studies used a binary assessment, the clinician had to choose between the occurrence or nonoccurrence of an event. This meant that there was no consideration of the clinicians' doubt—which is often the case in clinical practice—this might have underestimated the clinician's performance. The implementation of ordinal (such as, occurrence, doubt, or non-occurrence) or continuous (percentage of confidence of the occurrence) could mimic a more realistic environment in future comparative studies. Eighth, none of the studies adhered to the TRIPOD guideline, in particular the subitems of model specification and development. Following this statement is important to promote uniformity in presenting and developing ML models, thereby allowing future studies to be compared.[40] Lastly, no study included speed as a performance measure. In simple and repetitive tasks, the computer is increasingly expected to outperform humans on this measure

Machine learning models provided, in aggregate, only very slight improvements in diagnostic accuracy, sensitivity, and specificity compared with clinicians working alone. In the similar study by Senders et al.[24], we found an overall stronger performance of ML models compared with clinicians in neurosurgery. This might be explained by the fact that none of the included ML models in the current study used clinical input features such as age or vital parameters. The relationship between clinical parameters and outcomes such as postoperative survival is considerably more intricate, and especially in prognostication ML models may outperform clinicians. Several non-radiology orthopaedic ML models exist but none have been compared with humans to date.[6,52–54] In our earlier neurosurgery study, 10 of 23 studies compared ML models using clinical features as input with clinicians in predicting outcomes. All 10 demonstrated overall better performance of ML models compared with clinicians. Future studies should investigate the potential benefit of ML models using non-radiology input features to predict outcomes such as presurgical planning or survival in orthopaedic patients to determine the added value of these kind of algorithms.

ML models were primarily used to interpret radiologic data with the use of neural networks. Overall, ML models outperformed clinicians more when interpreting plain radiographs than when interpreting MRIs. Studies that investigated interpretation of plain radiographs looked at single radiographs showing osseous structures, while a series of MR images were converted to a two-dimensional (2D) image showing various structures. Additionally, the availability of training data for ML models that interpret plain radiographs is much higher than for ML models that interpret MRIs. This is reflected in the size of the datasets; plain radiographs had a larger median dataset than MRIs did: 2116 (IQR 1073-24,754) datapoints and 1370 (IQR 161-17,395) datapoints, respectively. As a recent study demonstrated, an increase in the size of training dataset to around 5000 images corresponded with increased performance, after which no benefit of additional training data was noticed.[55] Diversity in the predicted outcomes also influences on ML models' performance. In Chung et al.[41], distinctive fracture lines in the greater tuberosity with low variability made detection easier compared with fractures in the more complex anatomical surgical neck site. The same applies for detecting an osseous abnormality versus soft tissue abnormality–in general osseous abnormalities are more evident on imaging resulting in a better ML models' performance. Detection of "simple" osseous abnormalities on relatively uncomplicated plain radiographs might thus yield a higher difference in performance than complex MR images.

Radiologists and orthopaedists generally performed similarly to ML models, while ML models mostly outperformed other non-expert clinicians. This suggests that ML models can improve health care by assisting in well-defined tasks for non-musculoskeletal specialists or trainees and can aid clinicians in more austere or remote settings. Our neurosurgical review included studies that compared ML models and clinicians subdivided by specialty, but no separate analyses were provided to make a comparison.[24]

Considerable improvements were demonstrated in diagnostic accuracy of specialists aided by ML models. In orthopaedics, the potential benefit of lower misinterpretation rates of radiographs is especially worthwhile. In addition to potential liability issues[25], misdiagnosed radiographs may have severe clinical consequences such as joint collapse and posttraumatic osteoarthritis. Also, assessing abnormalities of the musculoskeletal system on imaging comprises a significant amount of time during daily orthopaedic practice. Clinicians face an increasing amount of imaging studies and complexity compared with 10 to 20 years ago, making it both time consuming and more prone to error.[56] Multiple studies suggest that time devoted to imaging interpretation decreases when aided by ML models compared with non-aided time.[31,48,57] This emphasizes that these ML models could improve the safety and effectiveness of patient care while working in conjunction with human counterparts.

We found that ML models have comparable performance to clinicians in assessing musculoskeletal images. ML models may enhance the performance of clinicians as a technical supplement rather than as a replacement for clinical or natural intelligence. On the other hand, there are circumstances in which ML models perform tasks that lie beyond the capacity of clinicians, such as accurately predicting complications and survival in patients with cancer.[16,17,20,53,58] Additionally, the advantages of using computers in helping make clinical decisions–such as uninterruptedly working at a high speed without fatigue–hold great potential to improve healthcare. Future studies should emphasize how ML models can complement clinicians, instead of analyzing the potential superiority of one versus the other. Substantial challenges exist before ML can be used regularly in daily practice. The sterile research environments in which algorithms are developed do not reflect the conditions observed in clinical practice. Also, ML models often reveal connections between disease characteristics and clinical outcomes in ways humans cannot understand.[59] This results in a lack of explanation or rationale for the crucial decisions ML models make, which is currently known as the "black box problem." Clinicians could be guided toward incorrect decisions if the algorithm is not well understood. The heat map proposed by Lindsey et al.[48], could provide a solution to this issue. This heat map is overlaid on the radiograph and highlights the model's calculated probability of a fracture – from yellow when the models is more confident to blue when less confidence – without making the binary decision of the bone being fractured or not.

# CONCLUSION

The optimal synergy between man and machine can be achieved by improving transparent reporting, diminishing bias, determining feasibility of application in the clinical setting, and appropriately considering conclusions. In the future, orthopaedics will likely embrace machine learning as a technical supplement rather than as a replacement for clinicians, creating a desirable synergy between "machine and man" rather than "machine versus man."

# REFERENCES

1. Jordan MI, Mitchell TM. **Machine learning: Trends, perspectives, and prospects.** *Science.* 2015;349(6245):255–260.

2. Ghahramani Z. **Probabilistic machine learning and artificial intelligence.** *Nature.* 2015;521(7553):452–459.

3. Mitchell TM. **Machine Learning.** Vol. 1. New York: McGraw-HillScience; 1997.

4. Obermeyer Z, Emanuel EJ. **Predicting the future - big data, machine learning, and clinical medicine.** *N Engl J Med.* 2016;375(13):1216–1219.

5. Esteva A, Kuprel B, Novoa RA, et al. **Dermatologist-level classification of skin cancer with deep neural networks.** *Nature.* 2017;542(7639):115–118.

6. Cabitza F, Locoro A, Banfi G. **Machine learning in orthopedics: a literature review.** *Front Bioeng Biotechnol.* 2018;6:75.

7. Rajkomar A, Dean J, Kohane I. **Machine learning in medicine.** *N Engl J Med.* 2019;380(14):1347–1358.

8. Deo RC. **Machine learning in medicine.** *Circulation.* 2015;132(20):1920–1930.

9. Mahadevan S. **Average reward Rrinforcement learning: foundations, algorithms, and empirical results.** *Mach Learn.* 1996;22(1):159–195.

10. Bayliss L, Jones LD. **The role of artificial intelligence and machine learning in predicting orthopaedic outcomes.** *Bone Joint J.* 2019;101(12):1476–1478.

11. Karhade AV, Thio QCBS, Ogink PT, et al. **Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis.** *Neurosurgery.* 2019;85(1):E83–E91.

12. Karhade AV, Thio Q, Ogink P, et al. **Development of machine learning algorithms for prediction of 5-year spinal chordoma survival.** *World Neurosurg.* 2018;119(0):e842–e847.

13. Merrill RK, Ferrandino RM, Hoffman R, et al. **Machine learning accurately predicts short-term outcomes following open reduction and internal fixation of ankle fractures.** *J Foot Ankle Surg.* 2019;58(3):410–416.

14. Karhade AV, Bongers MER, Groot OQ, et al. **Natural language processing for automated detection of incidental durotomy.** *Spine J.* 2020;20:695-700

15. Wyles CC, Tibbo ME, Fu S, et al. **Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty.** *J Bone Joint Surg Am.* 2019;101(21):1931–1938.

16. Bongers MER, Thio QCBS, Karhade AV, et al. **Does the SORG algorithm predict 5-year survival in patients with chondrosarcoma? An external validation.** *Clin Orthop Relat Res.* 2019;477(10):2296–2303.

17. Karhade AV, Ogink PT, Thio QCBS, et al. **Development of machine learning algorithms for prediction of prolonged opioid prescription after surgery for lumbar disc herniation.** *Spine J.* 2019:19;1764-1771

18. Karhade AV, Thio QCBS, Kuverji M, et al. **Prognostic value of serum alkaline phosphatase in spinal metastatic disease.** *Br J Cancer.* 2019;120(6):640–646.

19. Thio QCBS, Karhade AV, Bindels BJJ, et al. **Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease.** *Clin Orthop Relat Res.* 2020;478(2):322–333.

20. Karhade AV, Ahmed AK, Pennington Z, et al. **External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease.** *Spine J.* 2020;20(1):14–21.

21. Hendrickx LAM, Sobol GL, Langerhuizen DWG, et al. A machine learning algorithm to predict the probability of (occult) posterior malleolar fractures associated with tibial shaft fractures to guide "malleolus first" fixation. *J Orthop. Trauma*. 2020;34(3):131–138.

22. Karhade AV, Schwab JH, Bedair HS. Development of machine learning algorithms for prediction of sustained postoperative opioid prescriptions after total hip arthroplasty. *J Arthroplasty*. 2019:34(10);2272-2277

23. Thirukumaran CP, Zaman A, Rubery PT, et al. Natural language processing for the identification of surgical site infections in orthopaedics. *J Bone Joint Surg Am*. 2019;101(24):2167–2174.

24. Senders JT, Arnaout O, Karhade AV, et al. Natural and artificial intelligence in neurosurgery: a systematic review. *Neurosurgery*. 2018;83(2):181–192.

25. Berlin L. Defending the "missed" radiographic diagnosis. *AJR Am J Roentgenol*. 2001;176(2):317–322.

26. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318(22):2211.

27. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749–760.

28. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116–119.

29. Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med*. 2019;25(3):433–438.

30. Lee H, Yune S, Mansouri M, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng*. 2019;3(3):173–182.

31. Hollon TC, Pandian B, Adapa AR, et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat Med*. 2020;26(1):52–58.

32. Milea D, Najjar RP, Zhubo J, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med*. 2020;382(18):1687–1695.

33. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.

34. Slim K, Nini E, Forestier D, et al. Methodological index for non-randomized studies (minors): development and validation of a new instrument. *ANZ J Surg*. 2003;73(9):712–6.

35. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Med*. 2018;15(11):e1002699.

36. Bureau NJ, Destrempes F, Acid S, et al. Diagnostic accuracy of echo envelope statistical modeling compared to B-mode and power doppler ultrasound imaging in patients with clinically diagnosed lateral epicondylosis of the elbow. *J Ultrasound Med*. 2019:38;2631-2641

37. Gioftsos G, Grieve DW. The use of artificial neural networks to identify patients with chronic low-back pain conditions from patterns of sit-to-stand manoeuvres. *Clin Biomech. (Bristol, Avon)*. 1996;11(5):275–280.

38. Chee CG, Kim Y, Kang Y, et al. Performance of a deep learning algorithm in detecting osteonecrosis of the femoral head on digital radiography: a comparison with assessments by radiologists. *AJR Am J Roentgenol*. 2019:1–8.

39. Piraino DW, Amartur SC, Richmond BJ, et al. Application of an artificial neural network in radiographic diagnosis. *J Digit Imaging*. 1991;4(4):226–232.

40. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med*. 2015;13:1.

41. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop*. 2018;89(4):468–473.

42. Kim K, Kim S, Lee YH, et al. Performance of the deep convolutional neural network based magnetic resonance image scoring algorithm for differentiating between tuberculous and pyogenic spondylitis. *Sci Rep*. 2018;8(1):13124.

43. Liu F, Zhou Z, Samsonov A, et al. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. *Radiology*. 2018;289(1):160–169.

44. Xue Y, Zhang R, Deng Y, et al. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. *PLoS One*. 2017;12(6):e0178992.

45. Urakawa T, Tanaka Y, Goto S, et al. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiol*. 2019;48(2):239–244.

46. Adams M, Chen W, Holcdorf D, et al. Computer vs human: deep learning versus perceptual training for the detection of neck of femur fractures. *J Med Imaging Radiat Oncol*. 2019;63(1):27–32.

47. Gan K, Xu D, Lin Y, et al. Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. *Acta Orthop*. 2019:1–12.

48. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci USA*. 2018;115(45):11591–11596.

49. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop*. 2017;88(6):581–586.

50. Wang P, Berzin TM, Glissen Brown JR, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*. 2019;68(10):1813–1819.

51. Lin H, Li R, Liu Z, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *EClinicalMedicine*. 2019;9:52–59.

52. Gabriel RA, Sharma BS, Doan CN, et al. A predictive model for determining patients not requiring prolonged hospital length of stay after elective primary total hip arthroplasty. *Anesth Analg*. 2019;129(1):43–50.

53. Karhade AV, Thio QCBS, Ogink PT, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. *Neurosurgery*. 2019:1;85:671-681

54. Ramkumar PN, Navarro SM, Haeberle HS, et al. Development and validation of a machine learning algorithm after primary total hip arthroplasty: applications to length of stay and payment models. *J Arthroplasty*. 2019;34(4):632–637.

55. Varma M, Lu M, Gardner R, et al. Automated abnormality detection in lower extremity radiographs using deep learning. *Nat Mach Intell*. 2019;1(12):578–583.

56. Mirvis SE. Increasing workloads in radiology: Does it matter? *Appl Radiol*. 2013;42(5):6.

57. Gilbert FJ, Astley SM, Gillan MGC, et al. Single reading with computer-aided detection for screening mammography. *N Engl J Med*. 2008;359(16):1675–1684.

58. Thio QCBS, Karhade AV, Ogink PT, et al. **Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma?** *Clin Orthop Relat Res.* 2018;476(10):2040–2048.

59. Reardon S. **Rise of robot radiologists.** *Nature.* 2019;576(7787):S54–S58.

# SUPPLEMENTAL MATERIAL TO CHAPTER 19

**Appendix 1.** Search syntaxes for the PubMed, Embase, and Cochrane Library on October 1st, 2019

**Appendix 2.** Critical appraisal of 14 studies

**Appendix 3.** Included studies assessed according to Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD)

Supplemental material can be consulted online per the website of the journal and/or publisher.

## PART VI

# Summary and General Discussion

# CHAPTER 20

## SUMMARY

# SUMMARY

The number of patients with bone metastases is projected to increase within the next decade. Although non-operative treatment modalities for bone metastases have improved considerably over time, surgical treatment remains often indicated. Comprehensive understanding of the benefits and adverse events is essential in the pursuit of selecting the optimal candidate for surgical intervention and improving outcomes for these patients. This thesis aims to improve patient selection for surgical treatment of bone metastases by evaluating national trends (Part I) and quality-of-life benefits (Part II), identifying and predicting adverse events (Part III) with the help of Artificial Intelligence tools using patient and tumor characteristics (Part IV), and discussing challenges associated with Artificial Intelligence tools (Part V).

# PART I: RISING INCIDENCE

### Chapter 2. National Trends

The Nationwide Readmissions Database is an annual, multistate database that records approximately 50% million discharges each year in the United States. This national database provides national estimates across all age groups and diseases by weighting. Using these weighted, national estimates, the number of patients with bone metastases undergoing surgical treatment increased with 6.7% from 31,274 in 2016 to 33,361 in 2018. Furthermore, surgical bone metastases patients are becoming increasingly complex patients because of ageing and multiple comorbidities.

# PART II: QUALITY OF LIFE AND PHYSICAL FUNCTION

### Chapter 3. Minimal Clinically Important Difference

This prospective study determining the minimal clinically important difference (MCID) included 33 patients surgically treated for pathological fractures due to bone metastases in the lower extremity. The PROMIS MCIDs (95% confidence interval) for Pain Interference was 7.5 (3.4–12), Physical Function 4.1 (0.6–7.6), and Global Physical Health 4.2 (2.0–6.6); no MCID could be established for PROMIS Global Mental Health.

### Chapter 4. Meta-Analysis on Quality-of-Life Benefits in Spinal Metastases

In this meta-analysis of 10 studies, pooled data showed that in patients operated for spinal metastases with various indications including pain, spinal cord compression, instability, and tumor control, QoL rapidly improved and remained stable during the first 12 months after surgery. The pooled

QoL summary score improved from baseline to 1-month (standardized mean difference (SMD)=1.09, p<0.001), to 3-months (SMD=1.28, p<0.001), to 6-months (SMD=1.21, p<0.001), and to 9-12 months (SMD=1.08, p=0.001

### Chapter 5. Cohabitants Alternative Quality-of-Life Raters

This cross-sectional study included 47 patient-cohabitant pairs who independently completed QoL questionnaires for three PROMIS domains (anxiety, pain, and depression) with respect to the patients' symptoms. There were no MCIDs between the scores of patients and their cohabitants for all questionnaires, and the agreement between patient and cohabitant scores was moderate to strong (Spearman: 0.52 to 0.72). Despite the good agreement in QoL, the cohabitants' higher depression scores were correlated with increased differences in the anxiety and depression domains on the PROMIS.

## PART III: MORTALITY AND COMPLICATIONS

### Chapter 6. Impending Versus Completed Pathological Long Bone Fractures

After matching on 22 confounders, 270 impending pathological fractures were matched to 270 completed pathological fractures. Completed pathological fractures were defined as a destructive bone lesion with a visible fracture line, angulation, loss of height and/or rotation. Patients treated for an impending pathological fracture had better 1-year survival rate, less intraoperative blood loss, fewer perioperative blood transfusions, shorter anesthesia time, and fewer reoperations than patients treated for completed long bone pathological fractures. No differences were found for 30-days postoperative complications or hospitalization duration.

### Chapter 7. Venous Thromboembolism in Long Bone Metastases

This retrospective cohort study of 682 patients undergoing surgery for long bone metastases identified 6% (44/682) venous thromboembolisms within 90 days of surgery; 22 patients sustained a deep venous thromboembolism, and 22 a pulmonary embolism. The presence of venous thromboembolism resulted in a worse 1-year survival rate (27%) compared with non-venous thromboembolism (39%). No association was found between the use of chemoprophylaxis and venous thromboembolisms or wound complications.

### Chapter 8. Venous Thromboembolism in Spinal Metastases

In this retrospective cohort study of 637 patients, 11% (72/637) had symptomatic venous thromboembolisms; 6% (40/637) developed a deep venous thromboembolism and 6% (38/637) a pulmonary embolism of which 1.3% (8/637) were fatal. Patients with symptomatic venous thromboembolisms had a worse 1-year survival rate (38%) compared with non-venous thromboembolisms (47%). The overall proportion of patients that developed a wound complication is 10% (66/637), including 1.1% (7/637) spinal epidural hematomas. No association was found between any of the different chemoprophylaxis regimens and the development of symptomatic venous thromboembolisms or postoperative wound complications.

### Chapter 9. Adverse Events in Spinal Bone Metastases

Two affiliated tertiary institutions in Boston, the United States contributed to this retrospective cohort study of 647 patients with spinal metastases undergoing surgical treatment. From the 647 patients, 32% (205/647) had a 30-day complication rate, and 18% (115/647) had at least one reoperation. Complications within 30-days had a negative impact on survival (hazard ratio=1.63). Reoperations did not affect survival. The neurologic status remained equal in most patients, and surgery could improve the neurologic status in about 20% of all patients.

## PART IV: SUPPLEMENTING ARTIFICIAL INTELLIGENCE TOOLS

### Chapters 10, 11, and 12. CT Measurements as Predictor for Adverse Events

Chapters 10 and 11 included 212 patients with preoperative CT scans undergoing surgery for long bone metastases. Sarcopenia, defined as total muscle area measured on the level of the 4th lumbar vertebra using an in-house deep learning algorithm divided by the height squared, was a strong predictor for both 90-day and 1-year mortality. As for secondary outcomes, increased visceral abdominal tissue area was associated with increased length of stay and increased muscle area was associated with increased chance of reoperation. No body composition measurements were associated with postoperative complications within 30 days.

Chapter 12 included 196 patients with preoperative CT scans undergoing surgery for spinal metastases. Decreased muscle area was a predictor for development of postoperative complications within 30 days while controlling for confounding clinical factors. No body composition measurements were associated with length of stay and reoperations. Sarcopenia was a predictor for both 90-day and 1-year mortality (not included in this thesis).

### Chapter 13. External Validation of Prognostication Tool on American Dataset

A previously developed AI model that predicted survival in patients undergoing surgery for long bone metastases based on 15 clinical parameters including primary tumor, visceral metastases and systemic therapy had yet to be tested in external data. This external validation used 264 patients from the University of Iowa to assess the performance. The model retained good discriminative ability (area under the curve (AUC) 0.83 for 90-day mortality and AUC 0.84 for 1-year mortality), calibration, and decision curve analysis.

### Chapter 14. External Validation of Prognostication Tool on Asian Dataset

The model in Chapter 13 remained to be externally validated in an Asian cohort. Furthermore, the Eastern Cooperative Oncology Group (ECOG) performance status, a survival prognosticator repeatedly validated in other studies, was not considered into the algorithms because of missing data in the development cohort. The Taiwanese cohort had more patients with comorbidities, rapid primary tumor growth, ECOG score of 3 or 4, preoperative systemic therapy, preoperative local radiation, and less other bone metastases. Despite the baseline differences, the AI model generalized well in a Taiwanese cohort of 356 patients in terms of both discrimination and decision curve analysis. ECOG performance status provided additional prognostic value for 90-day mortality prediction.

### Chapter 15. Influence of Geographic Distinct Regions on Artificial Intelligence Models

Two AI models exist that predict survival in patients with spinal metastases based on various clinical parameters.ne difference is the inclusion of BMI by one model. This meta-analysis used four American and one Asian external validation cohort to demonstrate that both AI models, developed with American cohorts, performed better in American than in Asian patients. In addition, AI algorithms that did not incorporate demographic-specific variables such as BMI as input were less influenced in performance by non-American cohorts.

### Chapter 16. Natural Language Processing to Process Free-Text

A total of 704 bone scintigraphy reports of patients undergoing surgery for bone metastases were labeled each by three independent reviewers using a binary classification (single metastasis versus two or more metastases) to establish a ground truth. A stratified 80:20 split was used to develop and test a natural language processing (NLP) algorithm. The NLP algorithm correctly identified multiple bone metastases in 117 of the 124 who had multiple bone metastases in the testing cohort (sensitivity 0.94) and yielded 3 false positives (specificity 0.82).

# PART V: STRENGTHS AND LIMITATIONS OF ARTIFICIAL INTELLIGENCE

## Chapter 17. Overview of Machine Leaning Prognostication Tools in Orthopaedic Surgery

This review illustrated that 59 prediction models in orthopaedic surgery have been developed up to June 2020. Of the 59 models, 41% were at high risk of bias mainly because of incomplete reporting of performance measures. In addition, the overall median completeness of the Transparent Reporting of the multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) checklist was 53% (interquartile range 47%–60%). The abstract was only fully reported in 3% (2/59).

## Chapter 18. Overview of Externally Validated Prognostication Tools

Of the 59 ML models identified in Chapter 17, only 10 models (17%) were validated. All external validations identified in this review retained good discrimination. The external validation studies were characterized by incomplete reporting of performance measures, limiting a transparent examination of model performance. The overall median TRIPOD completeness was 61% (interquartile range 43%–89%).

## Chapter 19. Artificial Intelligence versus Clinicians in Interpreting Musculoskeletal Images

This review demonstrated that AI models showed, across 12 studies, slight improvements in diagnostic accuracy and sensitivity compared with clinicians working alone and were on par in specificity in interpreting musculoskeletal abnormalities (3% IQR-2.0% to 7.5%; 0.06% IQR -0.03 to 0.14; and 0.00 IQR -0.048 to 0.048, respectively). Orthopaedic surgeons and radiologists performed similarly to AI models, while AI models mostly outperformed other clinicians such as physiotherapists, general physicians, and emergency medicine clinicians (outperformance in 7 of 19, 7 of 23, and 6 of 10 performance comparisons, respectively). Two studies evaluated the performance of clinicians aided and unaided by AI. Both studies demonstrated considerable improvements in AI-aided clinician performance by reporting a 47% decrease of misinterpretation rate and a mean increase in specificity of 0.048.

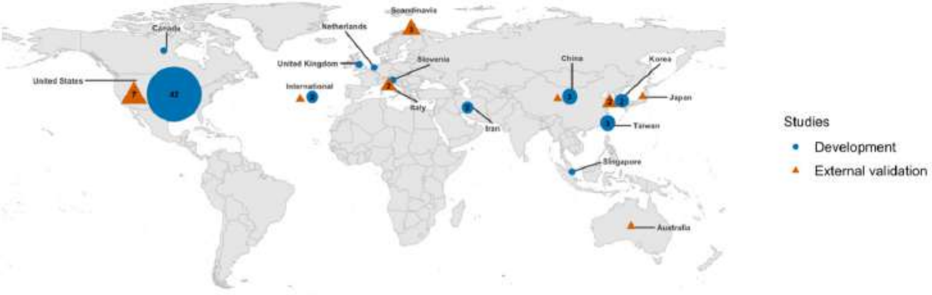


**Figure 1.** Worldwide distribution of development (n=59) and external validation studies (n=18).

CHAPTER

21

# GENERAL DISCUSSION AND FUTURE PERSPECTIVES

## PART I: RISING INCIDENCE

The number of patients with bone metastases who undergo surgical treatment is increasing. The stress this patient group places on the healthcare system is not limited to surgeons but extends (and is not limited) to emergency departments, physical therapists, and rehabilitation centers. Demand for this type of palliative surgery is growing, which highlights the need for increased preparation and optimal patient selection. Multidisciplinary approaches are needed to formulate individualized plans, using risk factors or prediction tools that have been developed in subpopulations based on patient and disease specific characteristics. The remaining chapters of this thesis work to improve the patient selection process for surgical treatment of bone metastases and the potentially beneficial role of AI.

## PART II: QUALITY OF LIFE AND PHYSICAL FUNCTION

Quality-of-Life (QoL) studies for patients with advanced diseases, such as bone metastases, are often relatively small with missing data, especially compared to other surgeries like hip replacements. This is due to high mortality rate, intensity of disease progression, and severity of symptoms.[2,3] Also, adequately measuring QoL is challenging because of the high number of patients lost to follow-up, the influence of comorbidities, and compounding psychosocial and emotional issues.[8–10] Both Chapters 3 and 4 experience low participation and completion rates (Chapter 3=51%; 33/65), and small sample sizes.[4,5] The small sample size in Chapter 3 limits a sub-analysis for whether MCID values or QoL benefits differ according to clinical or demographic characteristics. For example, completed pathological fractures are known to be associated with worse outcomes as compared with impending fractures (see: Chapter 6).[6,7] This may correspond with smaller MCIDs/QoL benefits as these patients experience less postoperative improvement in mobility and pain as compared with impending fractures. To clarify a more definite MCID/QoL benefit for each subpopulation, a larger cohort is required to provide stratified results by baseline characteristics. Until then, the MCIDs function as a benchmark to provide valuable information in managing expectations for clinicians and patients during treatment.

Although Chapter 4 suggests that surgery improves QoL, it is paramount to carefully select those patients who would benefit (most) from a surgical treatment. In addition, the QoL improvement may only be partially attributable to the operation, as the change in QoL may be mostly resulting from perioperative strategies. Radiotherapy and systematic therapy, for example, are not taken into account by the included studies. It is also important to consider potential adverse events, notably postoperative complications within thirty days, which were present in up to 35% of cases (see: Part III).[11] Nevertheless, these study results can be used to inform patients on postoperative expectations and help physicians to understand the potential postoperative course and use this for better decision-making. Future research should report clear definitions of selection criteria and surgical indication and provide stratified QoL results by indication and clinical characteristics such as primary tumor type, preoperative Karnofsky and Bilsky scores to elucidate the optimal candidate for surgical intervention.

The finding that cohabitants may be a reliable alternative to rate QoL instead of patients in Chapter 5 is important because it can address some of the limitations described in Chapter 3 and Chapter 4. Up to 70% of patients with advanced cancer are unable to complete QoL forms which accounts for the "missing data" rate in the two previous studies.[12] However, the cohabitant's mental and emotional condition is often adversely affected because they are fulfilling a demanding role in managing and supporting a patient with advanced cancer.[13–15] For example, if a cohabitant has depression, he/she is likely to overestimate a patient's symptoms in emotional domains, pushing clinicians to reconsider relying on judgments from cohabitants with signs of depression. Although cohabitants may be reliable alternatives to patients who are unable to complete QoL questionnaires, patient self-reported QoL remains preferred.

Part II underscores the importance of QoL outcomes in this fragile population. In general, two major limitations complicate QoL research in patients with bone metastases. First, high rates of missing data are present due to loss of follow-up. This lack of full data severely limits possible analysis and introduces bias as patients with severe symptoms or disease progression often do not complete QoL questionnaires.[12] Second, conducting QoL questionnaires are time-consuming and burdensome, often resulting in patient's abstention. However, excluding some patient responses leads to selection bias and smaller sample size. Both limitations could be resolved through using alternative raters, for example treating clinicians.[16–19] Yet, clinicians often vary considerably during treatment, while cohabitants interact with patients during an extended period and in a range of circumstances. Based on the findings in Chapter 5, the use of cohabitants for surveys should be recommended if a patient is unwilling or unable to personally complete QoL questionnaires.

## PART III: MORTALITY AND COMPLICATIONS

Patients undergoing surgery for an impending pathological fracture have a lower 1-year mortality rate and improved secondary outcomes compared with patients undergoing surgery for a completed pathological fracture. This chapter 6 supersedes prior work by using institutional data and thorough propensity score matching for 22 confounding factors.[21–28] This study design allows for the more evidenced conclusion that prophylactic surgery is superior to treating completed fractures.[29–31] Lesions that are causing disability and are at-risk for developing a completed fracture should ideally be correctly identified, which will prevent completed fractures and unnecessary surgical intervention. At present, predictive models for fractures are inaccurate and are not user-friendly. For example, the widely known Mirels score lacks sufficient specificity and sensitivity, and interobserver agreement is moderate as the components of the score are found to be subjective.[32,33] CT-based predictive algorithms show promising results, but clinical application is limited due to selection bias and challenging interface.[34,35] To benefit the clinical oncological practice, future research should focus on an accessible, easy-to-use, and accurate prediction tool that identifies a patient's risk level to develop a completed pathological fracture. With this tool, patients who may benefit from prophylactic surgical stabilization can be identified.

Due to the retrospective nature of Chapter 7, Chapter 8, and Chapter 9, the rates of adverse events are, although high, likely underestimated. Complications and reoperations after discharge have most likely been missed as patients often travel to specialized institutions like Massachusetts General Hospital to receive surgery. Because of this shift in location of care, it is difficult to keep track on all patients within the healthcare system. Follow-up consults are also often with home state hospitals, making tracking even more difficult. Additionally, the lack of a screening protocol at Massachusetts General Hospital could miss asymptomatic adverse events, especially important for complications like venous thromboembolism. Admittedly, it is possible that asymptomatic events are of less clinical relevance. Nevertheless, the risk for postoperative adverse events is high in patients with bone metastases undergoing surgical treatment, warranting deep consideration before initiation of intervention. Surgeons, together with patients, should be aware of the considerable risk of adverse events. Prognostic AI tools as addressed in Part IV should be developed for adverse events to help identify patients at high risk. At present, none of these personalized prediction tools exist. By providing adverse event predictions, patients and their physician can come to an informed, shared decision whether to opt for surgical treatment.

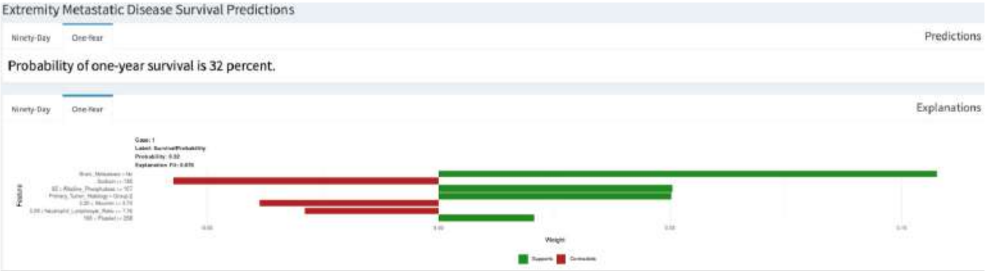
Part III contributes to a greater understanding of the prevalence and risk factors of adverse events, but clinicians need additional strategies to face the increasing incidence of bone metastases exacerbating the strain on the healthcare system.

## PART IV: SUPPLEMENTING ARTIFICIAL INTELLIGENCE TOOLS

Part IV presents AI-powered tools as a potential resource to improve patient selection and prognostication for surgical intervention. It is important to note that we did not deduce the benefits of adding CT scans to the existing prognostication models using clinical variables in Chapter 10, Chapter 11, and Chapter 12.[2,3,36,37] Sarcopenia and other CT measurements show a relation with survival and complications. Unfortunately, the number of included CT-scans was too low to explore added benefits. The relatively small sample size also explains the inability to control for known confounders.[38] Future multi-institutional collaboration should address these concerns and compare the value-add of CT characteristics to current standard prognostication tools like SORG, SINS, PATHFx, Bollen, and NESMS.[3,39–41] Yet, these three chapters suggest the potential of "hidden" prognostic parameters in CT scans that currently are largely ignored. CT scans are often performed preoperatively in patients with neoplastic diseases and are thus readily available to augment existing prognostication tools. AI algorithms should be integrated into the electronic healthcare system to automatically process CTs and extract meaningful CT parameters. The CT parameters in combination with clinical prognostication models can improve accurate and reliable survival prediction. These predictions can be used for shared decision making for patients with bone metastases that are considering surgical management

The previously developed AI model that predicts survival in patients undergoing surgery for long bone metastases continues to discriminate well in both Chapter 13 and Chapter 14. In Chapter 13, the validation cohort is highly homogeneous in disease traits and geographically similar to the development cohort, both from American tertiary hospitals.[2] The algorithm has not yet been tested in a predominantly non-American population. Treatment, patient and disease characteristics may be different in other demographic groupes.[42] For example, tumor biology has shown large variations by ethnicity, and access to care and quality of surgical cancer treatment may be distinct based on region.[43] Chapter 14 and Chapter 15 address these concerns.

The TRIPOD guideline encourages to repeatedly validate all AI algorithms, particularly in demographic distinct regions.[44] Chapter 14 validates the same model from Chapter 13 in a non-American cohort. Although the model retains excellent discriminatory ability and provides clinical benefits on decision curve analysis, the results from calibration and Brier score analyses indicate that model adjustment might be necessary for patients of Han Chinese descent because the models tend to underestimate patient survival in the Taiwanese validation cohort. Future studies should improve the predictive models by incorporating demographic-specific variables as suggested in Chapter 15. Although the two algorithms in Chapter 15 share multiple similar prognosticators, various parameters are included in the BMI-model in addition to BMI such as American Spinal Injury Association (ASIA) impairment scale and laboratory values (Figure 1).[34,45] These additional confounding factors can also explain the difference in the observed performance. Therefore, this meta-analysis can only indirectly support the argument that demographic-specific BMI can be more useful than BMI itself since confounding factors exist. In addition, the internal mechanism of the AI models is undisclosed and the pseudonymized data is unavailable.[42] Despite these limitations, future models should address the demographic-specific concerns highlighted in this meta-analysis. Existing prognostication models may need to recalibrate and optimize by considering demographic-specific variables such as BMI. This optimization requires an increased effort of international collaboration so more patients across demographic distinct regions can benefit from these promising prediction algorithms. Beside including demographic-specific variables, AI models need to keep improving by considering incremental factors such as ECOG, tumor mutation profiles, novel systemic therapies or body composition measurements based on imaging (see Chapter 10, 11, and 12). Future research would also benefit from determining the efficacy of this algorithm for non-operative management strategies. What is the utility in predicting survival, irrespective of selected treatment strategy?



Probability of one-year survival is 32 percent.

**Figure 1.** Patient X is considering surgical treatment for a femoral metastatic lesion at-risk for fracture and wants to know his survival expectancy. The AI model predicts a 1-year survival probability of 32%. The green bars visualize the variables from Patient X that favor survival: no brain metastases, alkaline phosphatase level between 82 and 107 IU/L, moderate-growth primary tumor, and platelet count between 193 and 258 x103/uL. The red bars are variables that result in an adjustment that increases the probability of mortality: sodium level higher than 136 mg/dL, albumin level between 3.3 and 3.7 g/dL, and neutrophil-to-lymphocyte ratio between 5.3 and 7.8. At the same time, consider that patient Y has a similar femoral metastatic lesion but has different variables that favor survival resulting in a 1-year survival probability of 85%. Depending on these survival predictions, patient X may be less inclined to pursue surgical treatment due to a lower survival rate compared with patient Y. In the future, models should be developed to predict outcomes for patients who choose to forgo operations. The clinical characteristics of each individual patient can be filled out online for free at https://sorg-apps.shinyapps.io/extremitymetssurvival/.

The most evident limitation of the NLP presented in chapter 16 is 'missing the point', or its inability to tease out specific language.[46] In a test, Microsoft ran twenty Shakespeare plays through NLP to map out emotions. While it deciphered extreme emotions well, it had trouble deciphering comic from tragic. In the words of a Microsoft developer, "The algorithm couldn't work out whether Hamlet's mad ravings were real or imagined, whether characters were being deceptive or telling the truth. That meant that the AI labeled events as positive when they were negative, and vice-versa. The AI believed The Comedy of Errors was a tragedy because of the physical, slapstick moments in the play."[47] This problem also applies to medical language that contains ambiguous, unintuitive, technical or abstract vocabulary that requires clinical interpretation.[48] In addition, the generalizability of NLP may not be transferable to different divisions or hospitals. External validation is needed before clinicians can use this tool in other medical institutions. After external validation, these NLP algorithms should be integrated into the electronic health care system to supplement procedural or diagnostic codes, and bypass error prone and labor-intensive manual chart review to extract meaningful clinical features.

Part IV demonstrates that AI tools have matured to become an important part of data-driven healthcare. In parallel, clinicians are increasingly appreciating the added value of AI. Yet, while patient reported outcomes measures are considered to be the most important outcome in this population, most AI studies focus on survival and adverse surgical events. Future multicenter efforts should seek to create registries of patients with both sufficient volume and quality of life data to realize reliable AI models. Cohabitants can be used to ensure complete quality-of-life data (Chapter 5).

More pressing, no validated prediction models exist for patients with bone metastases that choose for active monitoring or non-surgical treatments such as radiotherapy, systemic treatment, or a combination thereof. For example, a patient who presents with a femoral lesion at risk of breaking is considering surgery, active monitoring, or radiotherapy. With the available models, only a survival prediction can be made for the surgical option. No (AI) prognostic tools exist for non-surgical options that can aid the decision-making process for both patients and physicians. To benefit the clinical oncological practice, future research should develop easy-to-use and accurate prediction tools that identify a patient's benefit level across a range of treatment strategies.

## PART V: STRENGTHS AND LIMITATIONS OF ARTIFICIAL INTELLIGENCE

Despite the evident potential of AI, the findings in Chapters 17, 18, and 19 lead to several challenges that need to be addressed before AI algorithms can be accepted in daily practice (Figure 2). First, model development and transparent reporting directions should be followed beginning in the early stages of study design. The TRIPOD checklist can be used to ensure this. Transparent and complete reporting is required to critically assess the presence of bias, facilitate study replication, and correctly interpret study results. Second, the development of clinical prediction models must be coupled with the intention of using them in clinical practice. Prior to utilization, validation of AI models on large, prospective, geographically distinct datasets is required. Only well-executed validation studies can ensure accurate and reliable AI models. Currently, only 17% of the available prognostic ML models in orthopaedic surgery are validated. Third, fewer than half of the published studies offer the means to calculate predictions through web calculators or in-study formulas, making external validation and individual predictions difficult.[49] Ideally, the algorithms are published online to facilitate sharing and collaboration. Last, the majority of studies focuses solely on the stand-alone performance of AI models. Taking all of this into consideration, while AI enhances the performance of clinicians, it should be used as a technical supplement rather than as a replacement for natural intelligence.[50] Future studies should focus on this dichotomy.

In many cases of bone metastases, there will be clinical ambiguity in terms of the optimal treatment strategy, including surgical strategies. Surgical management of patients with bone metastases is, despite potential maintenance or improvement of QoL, resource-intensive and carries significant rates of perioperative morbidity. This thesis provides important aspects that should be considered before choosing the surgical pathway for patients with bone metastases. In all contexts, surgeons, together with patients, must weigh the likelihood of both benefits and adverse events when choosing a surgical intervention.

Patients with bone metastases considering treatment exemplify a challenge central to today's healthcare: there is a limit to what clinicians can accomplish. As patients with bone metastases continue to become more common, it is unlikely that clinicians can readily meet this higher demand. To help this disparity, AI can supplement clinicians in processing large amounts of data and aid in personalized decision-making. Yet, clinicians are sensitive to the idea that their accrued knowledge and ability will be hostilely replaced by machine labor. The default adversarial attitude of humans at odds with machines dates to the apocryphal tale of John Henry versus the steam engine. John Henry, an African American steel driver, challenged a steam-powered drill from the Industrial Revolution in the 1800s to a railway steel-driving race to protect his job. John Henry won, only to tragically die of exhaustion. This competition ended in Henry's death and machines supplanting human effort in industry. However, AI usage serving as a supplement to human diagnostics capabilities and expertise can further the paradigm shift for a cooperative relationship. Instead of feeling threatened, clinicians should embrace AI as a supplement by creating a desirable synergy between "human and machine" rather than the ancient opposition of "human versus machine."

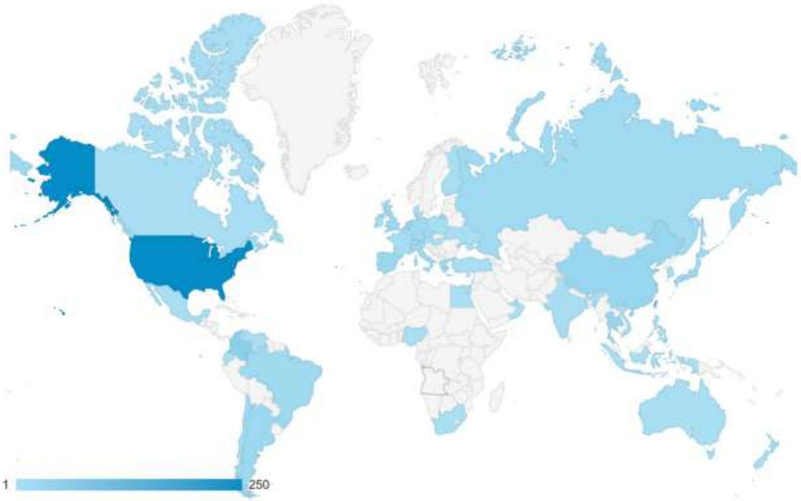


**Figure 2.** World heat map of the usage of the SORG prognostication model of spinal metastases in one year.

# MAIN STUDY QUESTIONS AND CONCLUSIONS

## Part I: Rising Incidence

Chapter 2 *Is there a growing trend in patients with bone metastases undergoing surgical treatment?*
The number of surgical bone metastases cases increased from 31,274 in 2016 to 33,361 in 2018, representing a 6.7% increase in the United States.

## Part II: Quality of Life and Physical Function

Chapter 3 *What are the MCID values of three PROMIS questionnaires in patients surgically treated for long bone metastases?*
The MCID values are for PROMIS Pain Interference 7.5 (3.4–12), PROMIS Physical Function 4.1 (0.6–7.6), and Global Physical Health 4.2 (2.0–6.6), helping set expectations of QoL benefits for the patient and clinician.

Chapter 4 *To what extent does surgery improve the QoL for patients with spinal metastases?*
In carefully selected patients with spinal metastases, surgery improves overall QoL and rapidly increased physical, emotional, and functional well-being; it has minimal effect on social/family well-being.

Chapter 5 *Do cohabitants reliably complete QoL questionnaires for patients with bone metastases?*
Cohabitants may be a reliable alternative to rate QoL in patients with bone metastases; this is potentially helpful in situations where the patient cannot weigh in.

## Part III: Mortality and Complications

Chapter 6 *What are the clinical outcome differences in the treatment of impending versus completed pathological long bone fractures?*
Patients undergoing surgery for impending pathological fractures have lower 1-year mortality rates and better secondary outcomes as compared with patients undergoing surgery for completed pathological fractures, while accounting for 22 confounders through propensity matching.

Chapter 7 *What is the incidence and impact on survival of VTE in long bone metastases?*
The rate of 30-day postoperative symptomatic VTE is 6% and the presence of VTE results in a worse 1-year survival rate compared with non-VTE patients.

Chapter 8 *What is the incidence and impact on survival of VTE in spinal metastases?*
The rate of 30-day postoperative symptomatic VTE is 11% – of which eight (1.3%) were fatal PEs – and the presence of VTE results in a worse 1-year survival rate compared with non-VTE patients.

Chapter 9 *What is the incidence and impact on survival of postoperative complications and reoperations in spinal metastases?*
The 30-day complication rate is 32% and reoperation rate is 18%. Complications within 30-days had a negative impact on survival; reoperations did not.

## Part IV: Supplementing Artificial Intelligence Tools

Chapter 10 *Can CT body compositions predict mortality in long bone metastases?*
Sarcopenia on level L4 is predictive of 90-day and 1-year mortality.

Chapter 11 *Can CT body compositions predict length of stay, complications, and reoperations in long bone metastases?*
Increased VAT area is associated with increased *length of stay*, increased muscle area is associated with increased chance of reoperation, and no associations are found with postoperative complications within 30 days.

Chapter 12 *Can CT body compositions predict length of stay, complications, and reoperation in spinal metastases?*
Decreased muscle area is a predictor for development of postoperative complications within 30 days; no body composition measurements are associated with length of stay and reoperations.

Chapter 13 *Can an AI algorithm accurately predict 90-day and 1-year mortality in an American external cohort of long bone metastases?*
The model retained good performance measures, lending support to the use of this AI tool in supplementing the clinical decision-making progress.

Chapter 14 *Can an AI algorithm accurately predict 90-day and 1-year mortality in an external cohort of long bone metastases?*
The model generalized well in a Taiwanese cohort in terms of both discrimination and decision curve analysis and the ECOG score was identified as an additional predictor for 90-day mortality.

Chapter 15 *Are AI models that do not incorporate demographic-specific variables less influenced in performance by geographic distinct regions?*

AI algorithms that do not incorporates demographic-specific variables such as BMI as input seem to be less influenced in performance by non-American cohorts.

Chapter 16 *Can an NLP algorithm automatically extract from radiology reports meaningful preoperative clinical variables?*

An AI-based NLP method automates the transformation of free text to binary classification of single and multiple bone metastases, thereby optimizing the speed, accuracy, and consistency of clinical chart review.

## Part V: Strengths and Limitations of Artificial Intelligence

Chapter 17 *What is the quality and transparent reporting of ML prognostic models in orthopaedic surgery?*

Of the 59 models that have been developed up to June 2020, 41% are at high risk of bias and over half incompletely reported their methods and/or performance measures.

Chapter 18 *How many ML models are externally validated?*

Most current predictive ML models are not externally validated: 18 studies externally validated 10 of the 59 models (17%).

Chapter 19 *Does AI outperform clinicians in interpreting musculoskeletal images?*

AI models have comparable performance to clinicians; however, AI improved the performance and speed of diagnosis when used as a supplemental tool.

# REFERENCES

1. Siegel RL, Miller KD, Fuchs HE, et al. Cancer statistics, 2021. *CA Cancer J Clin.* 2021;71(1):7–33.

2. Thio QCBS, Karhade AV, Ogink PT, et al. Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease. *Clin Orthop Relat Res.* 2020;478(2):322–333.

3. Karhade AV, Thio QCBS, Ogink PT, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. *Neurosurgery*. 2019:1;85:671-681

4. Zuckerman SL, Chotai S, Devin CJ, et al. Surgical resection of intradural extramedullary spinal tumors: Patient reported outcomes and minimum clinically important difference. *Spine (Phila. Pa. 1976).* 2016;41(24):1925–1932.

5. Wong E, Zhang L, Kerba M, et al. Minimal clinically important differences in the EORTC QLQ-BN20 in patients with brain metastases. *Support. Care Cancer*. 2015;23(9):2731–2737.

6. Harrington KD. New trends in the management of lower extremity metastases. *Clin Orthop Relat Res.* 1982;(169):53–61.

7. Francis KC. Prophylactic Internal Fixation of Metastatic Osseous Lesions. *Cancer*. 1960;13.

8. Ernst E, Filshie J, Hardy J. Evidence-based complementary medicine for palliative cancer care: Does it make sense? *Palliat Med.* 2003;17(8):704–707.

9. Choi D, Morris S, Crockard A, et al. Assessment of quality of life after surgery for spinal metastases: position statement of the Global Spine Tumour Study Group. *World Neurosurg.* 2013;80(6):e175-9.

10. Cheng EY. Prospective quality of life research in bony metastatic disease. *Clin Orthop Relat Res.* 2003;(415 Suppl):S289-97.

11. Fehlings MG, Nater A, Tetreault L, et al. Survival and clinical outcomes in surgically treated patients with metastatic epidural spinal cord compression: results of the prospective multicenter AOSpine study. *J Clin Oncol.* 2016;34(3):268–276.

12. Jones JM, McPherson CJ, Zimmermann C, et al. Assessing agreement between terminally ill cancer patients' reports of their quality of life and family caregiver and palliative care physician proxy Ratings. *J Pain Symptom Manage.* 2011;42(3):354–365.

13. Ji J, Zöller B, Sundquist K, et al. Increased risks of coronary heart disease and stroke among spousal caregivers of cancer patients. *Circulation*. 2012;125(14):1742–7.

14. O'Brien J, Francis A. The use of next-of-kin to estimate pain in cancer patients. *Pain*. 1988;35(2):171–8.

15. Vitaliano PP, Zhang J, Scanlan JM. Is caregiving hazardous to one's physical health? A meta-analysis. *Psychol Bull.* 2003;129(6):946–72.

16. Blazeby JM, Williams MH, Alderson D, et al. Observer variation in assessment of quality of life in patients with oesophageal cancer. *Br J Surg.* 1995;82(9):1200–3.

17. Sneeuw KCA, Aaronson NK, Sprangers MA, et al. Value of caregiver ratings in evaluating the quality of life of patients with cancer. *J Clin Oncol.* 1997;15(3):1206–17.

18. Sneeuw KCA, Aaronson NK, Sprangers MAG, et al. Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. *J Clin Epidemiol.* 1998;51(7):617–631.

19. Wilson KA, Dowling AJ, Abdolell M, et al. Perception of quality of life by patients, partners and treating physicians. *Qual Life Res.* 2000;9(9):1041–52.

20. Verlaan JJ, Choi D, Versteeg A, et al. Characteristics of patients who survived 3 months or 2 years after surgery for spinal metastases: can we avoid inappropriate patient selection? *J Clin Oncol.* 2016;34(25):3054–61.

21. Aneja A, Jiang JJ, Cohen-Rosenblum A, et al. Thromboembolic disease in patients with metastatic femoral lesions: a comparison between prophylactic fixation and fracture fixation. *J Bone Joint Surg Am.* 2017;99(4):315–323.

22. Philipp TC, Mikula JD, Doung Y-C, et al. Is There an Association Between Prophylactic Femur Stabilization and Survival in Patients with Metastatic Bone Disease? *Clin Orthop Relat Res.* 2020;478(3):540–546.

23. McLynn RP, Ondeck NT, Grauer JN, et al. What is the adverse event profile after prophylactic treatment of femoral shaft or distal femur metastases? *Clin Orthop Relat Res.* 2018;476(12):2381–2388.

24. El Abiad JM, Raad M, Puvanesarajah V, et al. Prophylactic versus postfracture stabilization for metastatic lesions of the long bones: A comparison of 30-day postoperative outcomes. *J Am Acad Orthop Surg.* 2019;27(15):e709–e716.

25. Blank AT, Lerman DM, Patel NM, et al. Is prophylactic intervention more cost-effective than the treatment of pathologic fractures in metastatic bone disease? *Clin Orthop Relat Res.* 2016;474(7):1563–1570.

26. Arvinius C, Parra JLC, Mateo LS, et al. Benefits of early intramedullary nailing in femoral metastases. *Int Orthop.* 2014;38(1):129–132.

27. Ristevski B, Jenkinson RJ, Stephen DJG, et al. Mortality and complications following stabilization of femoral metastatic lesions: a population-based study of regional variation and outcome. *Can J Surg.* 2009;52(4):302–308.

28. Ward WG, Holsenbeck S, Dorey FJ, et al. Metastatic disease of the femur: surgical treatment. *Clin Orthop Relat Res.* 2003;(415 Suppl):S230-44.

29. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med.* 2009;28(25):3083–3107.

30. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med.* 2014;33(7):1242–1258.

31. Austin PC, Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study. *Stat Methods Med Res.* 2016;25(5):2214–2237.

32. Mirels H. Metastatic disease in long bones. A proposed scoring system for diagnosing impending pathologic fractures. *Clin Orthop Relat Res.* 1989;(249):256–264.

33. Damron TA, Morgan H, Prakash D, et al. Critical evaluation of Mirels' rating system for impending pathologic fractures. *Clin Orthop Relat Res.* 2003;(415 Suppl):S201-7.

34. Janssen SJ, Paulino Pereira NR, Meijs TA, et al. Predicting pathological fracture in femoral metastases using a clinical CT scan based algorithm: A case-control study. *J Orthop. Sci.* 2018;23(2):394–402.

35. Snyder BD, Hauser-Kara DA, Hipp JA, et al. Predicting fracture through benign skeletal lesions with quantitative computed tomography. *J Bone Joint Surg. Am.* 2006;88(1):55–70.

36. Paulino Pereira NR, Janssen SJ, van Dijk E, et al. Development of a prognostic survival algorithm for patients with metastatic spine disease. *J Bone Jt Surg Am.* 2016;98(21):1767–1776.

37. Kapoor ND, Twining PK, Groot OQ, et al. Adipose tissue density on CT as a prognostic factor in patients with cancer: a systematic review. *Acta Oncol.* 2020;59(12):1488–1495.

38. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med.* 2019;38(7):1276–1296.

39. Forsberg JA, Wedin R, Bauer HCF, et al. External validation of the Bayesian Estimated Tools for Survival (BETS) models in patients with surgically treated skeletal metastases. *BMC Cancer*. 2012;12:493.

40. Schoenfeld AJ, Ferrone ML, Schwab JH, et al. Prospective validation of a clinical prediction score for survival in patients with spinal metastases: The New England Spinal Metastasis Score. *Spine J.* 2020;21:28-36

41. Fourney DR, Frangou EM, Ryken TC, et al. Spinal instability neoplastic score: An analysis of reliability and validity from the Spine Oncology Study Group. *J Clin Oncol.* 2011;29(22):3072–3077.

42. Yang JJ, Chen C-W, Fourman MS, et al. International external validation of the SORG machine learning algorithms for predicting 90-day and 1-year survival of patients with spine metastases using a Taiwanese cohort. *Spine J.* 2021:2;1529-9430

43. Ward E, Jemal A, Cokkinides V, et al. Cancer disparities by race/ethnicity and socioeconomic status. *CA Cancer J Clin.* 2004;54(2):78–93.

44. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* 2015;13:1.

45. Paulino Pereira NR, Janssen SJ, van Dijk E, et al. Development of a Prognostic Survival Algorithm for Patients with Metastatic Spine Disease. *J Bone Jt Surg.* 2016;98(21):1767–1776.

46. Hofstadter D. The shallowness of google translate. *Atl.* 2018;30.

47. Dashika Gnaneswaran. Breaking bard: using microsoft AI to unlock Shakespeare's greatest works. *Microsoft.* 2019

48. International Business Machines. IBM's Watson computer takes the Jeopardy! Challenge.

49. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol.* 2014;14:40.

50. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci USA.* 2018;115(45):11591–11596.

# APPENDICES

# NEDERLANDSTALIGE SAMENVATTING

Het aantal patiënten met botmetastasen neemt naar verwachting aanzienlijk toe in het komende decennium. Hoewel de mogelijke behandelingen voor botmetastasen aanzienlijk zijn verbeterd, blijft chirurgisch ingrijpen vaak aangewezen. Patiënten met botmetastasen zijn gebaat bij een nadere analyse van de voor- en nadelen van chirurgische behandeling. Dit proefschrift wil bijdragen aan een betere selectie van patiënten voor chirurgische behandeling van botmetastasen. De thesis omvat vijf delen: de evaluatie van de nationale trends (Deel I) en van de onderzoeksresultaten naar de kwaliteit van leven (Deel II), de determinatie van complicaties (Deel III) en de voorspelling ervan met behulp van AI-tools (Deel IV). Tot slot bespreken we de uitdagingen van diezelfde AI-tools (Deel V).

# DEEL I: STIJGENDE INCIDENTIE

### Hoofdstuk 2. Nationale trends

Ontwikkelingen in de behandeling van neoplastische ziekte heeft het leven van veel patiënten verlengd, maar als neveneffect zien we een toenemende incidentie van botmetastasen. Deze studie op populatieniveau van de Verenigde Staten toont aan dat het aantal patiënten met botmetastasen dat een chirurgische behandeling ondergaat, is gestegen met 6,7% van 31,274 in 2016 naar 33,361 in 2018. Chirurgische botmetastasen -patiënten worden steeds complexer vanwege vergrijzing en comorbiditeiten. De slechte algemene gezondheidstoestand, in combinatie met een invasieve operatie, leidt tot schadelijke gevolgen van ziekenhuisopname en vergroot de kans op heropname. Deze bevindingen benadrukken het belang van een gepersonaliseerde kosten-batenanalyse die voor elke individuele patiënt moet worden toegepast als een chirurgische behandeling wordt overwogen.

# DEEL II: KWALITEIT VAN LEVEN EN FYSIEKE FUNCTIE

### Hoofdstuk 3: Minimaal klinisch relevant verschil

De klinische relevantie van veranderingen in kwaliteit van leven (QoL) na een operatie is vaak onduidelijk. Deze prospectieve studie bepaalt het minimale klinisch relevante verschil (MCID) voor Patient-Reported Outcomes Measurement Information System (PROMIS) Physical Function (4.1), PROMIS Pain Interference (7.5) en Global Physical Health (4.2) bij patiënten die operaties ondergingen voor pathologische fracturen vanwege botmetastasen in de onderste extremiteiten. De MCID-waarden helpen bij het vormen van QoL-verwachtingspatronen voor de patiënt, het gezin en het behandelend team.

### Hoofdstuk 4. Meta-analyse van de QoL-voordelen bij spinale metastasen

In deze meta-analyse van 10 studies, verbetert de gepoolde QoL-score van baseline tot 1 maand (standardized mean difference (SMD)=1,09, p <0,001), tot 3 maanden (SMD=1,28, p <0,001), tot 6 maanden (SMD=1,21, p <0,001) en tot 9-12 maanden (SMD=1,08, p=0,001). Deze QoL-voordelen kunnen alleen worden bereikt door eerst zorgvuldig patiënten te selecteren die baat hebben bij een operatie.

### Hoofdstuk 5. Partner alternatieve QoL-beoordelaar

Partners kunnen een betrouwbaar alternatief zijn om QoL te beoordelen bij patiënten met botmetastasen. In deze cross-sectionele studie hebben 47 patiënt-partners onafhankelijk QoL-vragenlijsten ingevuld over de symptomen van de patiënt in drie domeinen (angst, pijn en depressie). MCID's tussen de scores van patiënten en hun partners ontbreken voor de drie domeinen en de overeenkomst tussen de QoL-beoordeling van patiënten en partners is matig tot sterk (Spearman: 0.52 tot 0.72). Deze bevindingen suggereren dat een partner de QoL van een patiënt juist kan inschatten. Dit is nuttige informatie in situaties waarin de mening van patiënt ontbreekt.

## DEEL III: MORTALITEIT EN COMPLICATIES

### Hoofdstuk 6. Dreigende versus voltooide pathologische fracturen

De verschillen in klinische uitkomsten zijn niet goed vastgesteld bij patiënten die een operatie ondergaan voor dreigende versus reeds opgetreden pathologische botbreuken in de lange pijpbeenderen. Deze propensiteit score studie met 22 confounders toont aan dat patiënten die behandeld worden voor een dreigende pathologische fractuur betere klinische uitkomsten hebben dan voor reeds opgetreden pathologische botfracturen. Het gaat dan om een betere 1 jaar overleving, minder intra-operatief bloedverlies, minder perioperatieve bloedtransfusies, een kortere anesthesietijd en minder heroperaties. De ontwikkeling van een eenvoudig, nauwkeurig en gevalideerd voorspellingsinstrument is van belang om te bepalen of een patiënt met een botmetastasen het risico loopt op een pathologische fractuur.

### Hoofdstuk 7. Veneuze trombo-embolie bij botmetastasen in de pijpbeenderen

De combinatie van gemetastaseerd kanker en orthopedische chirurgie is in theorie extra risicovol voor een veneuze trombo-embolie (VTE). Deze retrospectieve cohortstudie van 682 patiënten die een operatie ondergingen voor botmetastasen in de pijpbeenderen, identificeert 6% (44/682) VTEs binnen 90 dagen na de operatie, 3% (22/682) ontwikkelt een diepe veneuze trombo-embolie en 3% (22/682) een longembolie. De aanwezigheid van VTE resulteert in een slechtere 1 jaar overlevingskans

(27%) vergeleken met patiënten die geen VTE hebben (39%).

### Hoofdstuk 8. Veneuze trombo-embolie bij spinale metastasen

In deze retrospectieve cohortstudie van 637 patiënten heeft 11% (72/637) een symptomatische VTE, 6% (40/637) ontwikkelt een diepe veneuze trombo-embolie en 6% (38/637) een longembolie. Patiënten met symptomatische VTE hebben een slechtere 1 jaar overlevingskans (38%) vergeleken met niet-VTE patiënten (47%). Uit Hoofdstukken 7 en 8 blijkt dat het risico op symptomatische VTE hoog is bij patiënten met botmetastasen die een chirurgische behandeling ondergaan. Verder onderzoek is nodig om preventiestrategieën te bepalen voor VTE-complicaties.

### Hoofdstuk 9. Complicaties bij spinale botmetastasen

De incidentie, risicofactoren en impact op de overleving van postoperatieve complicaties en heroperaties zijn niet goed vastgesteld bij patiënten die een operatie ondergaan voor spinale metastasen. Deze retrospectieve cohortstudie van 647 patiënten heeft 32% (205/647) complicaties binnen 30 dagen, en 18% (115/647) van de patiënten heeft ten minste één heroperatie ondergaan. Complicaties binnen 30 dagen hebben een negatieve invloed op de overleving (hazard ratio=1.63). Heroperaties hebben geen invloed op de overleving. Chirurgen en patiënten zijn zich nog onvoldoende bewust van het hoge percentage complicaties en heroperaties. Ook de geïdentificeerde risicofactoren verdienen meer aandacht bij het overwegen van een chirurgische behandeling.

## DEEL IV: ONDERSTEUNDE KUNSTMATIGE INTELLIGENTIE TOOLS

### Hoofdstukken 10, 11 en 12. CT-metingen als voorspeller voor mortaliteit en complicaties

Metingen van lichaamssamenstellingen door computertomografie (CT) kunnen dienen als biomarker voor de voorspelling van mortaliteit en complicaties bij patiënten met botmetastasen. Een intern AI-algoritme automatiseerde de metingen van oppervlakte en dichtheid van onderhuids vetweefsel, visceraal vetweefsel en spieren op het niveau van L4.

**Hoofdstukken 10 en 11** includeren 212 patiënten met preoperatieve CT-scans die een operatie ondergaan voor botmetastasen in de pijpbeenderen. Sarcopenie is een sterke voorspeller voor zowel 90 dagen als 1 jaar mortaliteit. Wat complicaties betreft, is meer visceraal vetweefsel geassocieerd met een langere hospitalisatie en is meer spieroppervlak geassocieerd met een grotere kans op heroperatie. Geen van de lichaamssamenstellingen zijn voorspellend voor postoperatieve complicaties binnen 30 dagen.

**Hoofdstuk 12** includeert 196 patiënten met preoperatieve CT-scans die een operatie ondergaan voor spinale metastasen. Een verminderd spieroppervlak is een voorspeller voor de ontwikkeling van postoperatieve complicaties binnen 30 dagen. Geen van de lichaamssamenstellingen zijn voorspellend voor hospitalisatie of heroperaties.

De geïdentificeerde CT-parameters van de lichaamssamenstellingen kunnen worden gebruikt als nieuwe beeldvormende biomarkers voor voorspelling van mortaliteit en complicaties bij patiënten die een abdominale CT ondergaan voor stadiëring of beoordeling van therapie. AI-algoritmen maken automatische verwerking van CT-scans mogelijk. Deze zijn vaak direct beschikbaar bij patiënten met vergevorderde kanker.

### Hoofdstuk 13. Externe validatie van voorspelmodel in Amerikaanse patiënten

Eerder is een AI-model ontwikkeld dat de overleving voorspelt bij patiënten die een operatie ondergaan voor botmetastasen in de pijpbeenderen. Dit model moet nog wel worden getest in een externe dataset. Deze externe validatie maakt gebruik van 264 patiënten van de Universiteit van Iowa om de prestaties te evalueren. Het model behoudt een goed onderscheidend vermogen (oppervlakte onder de curve (AUC) 0.83 voor 90-dagen mortaliteit en AUC 0.84 voor 1-jaars mortaliteit), kalibratie en analyse van de beslissingscurve. De validatieresultaten ondersteunen het gebruik van deze AI-tool als aanvulling op de klinische besluitvorming.

### Hoofdstuk 14. Externe validatie van voorspelmodel in Aziatische patiënten

Het model in Hoofdstuk 13 moet nog extern worden gevalideerd in een Aziatisch cohort en we hebben de Eastern Cooperative Oncology Group (ECOG) score, een overleving predictor die herhaaldelijk in andere onderzoeken is gevalideerd, niet in onze algoritmen meegenomen omwille van missende data. Het model behoudt goede resultaten in een Taiwanese cohort van 356 patiënten voor zowel discriminatie als analyse van de beslissingscurve. De ECOG-score leverde aanvullende voorspelwaarde op voor de 90-dagen mortaliteit.

### Hoofdstuk 15. Invloed van raciaal verschillende regio's op AI-modellen

Deze meta-analyse maakt gebruik van vier Amerikaanse cohorten en één Aziatisch cohort om aan te tonen dat AI-modellen, ontwikkeld met Amerikaanse cohorten, beter presteren in Amerikaanse dan in Aziatische patiënten. AI-algoritmen die geen raciaal specifieke variabelen zoals BMI als input gebruiken, worden minder beïnvloed in hun resultaten door niet-Amerikaanse cohorten. Dit benadrukt het belang van het opnemen van regio specifieke variabelen in bestaande prognostische modellen om ze generaliseerbaar te maken naar raciaal verschillende regio's.

### Hoofdstuk 16. Natural language processing om vrije tekst te analyseren

Het wijdverbreide gebruik van elektronische gezondheidsgegevens heeft geleid tot ongekende mogelijkheden voor het extraheren van klinische kenmerken uit medische notities. Deze studie toont aan dat natural language processsing (NLP) relevante klinische variabelen kan extraheren, zoals het aantal preoperatieve botmetastasen in een patiënt. Het potentieel van deze NLP-tool is aanzienlijk omdat het automatisch en nauwkeurig grote hoeveelheden patiënten data kan verwerken. Deze werkwijze omzeilt foutgevoelige en tijdrovende handmatige arbeid.

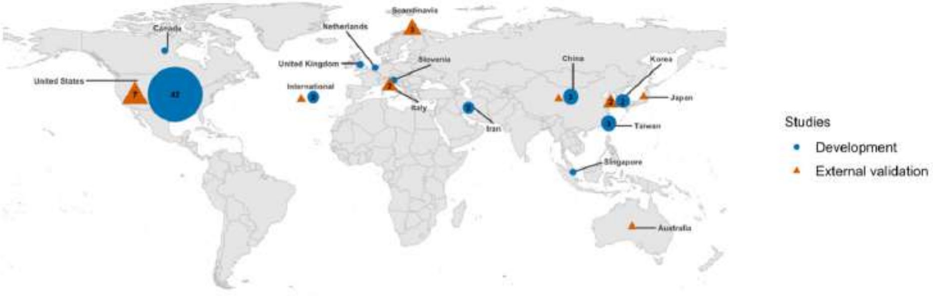
# DEEL V: VOOR- EN NADELEN VAN KUNSTMATIGE INTELLIGENTIE

### Hoofdstuk 17. Overzicht van machine-learning prognosetools in de orthopedische chirurgie

In de orthopedische chirurgie circuleert een groot aantal prognostische modellen op basis van machine learning (ML). Deze review illustreert dat 59 modellen zijn ontwikkeld tot juni 2020. Van de 59 modellen scoort 41% een hoog risico op bias, voornamelijk vanwege onvolledige rapportages van resultaten. De mediaan van de Transparent Reporting of the multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) is 53% (interkwartielafstand 47% -60%). Het abstract wordt slechts volledig gerapporteerd in 3% (2/59) van de studies. Deze bevindingen benadrukken het belang van methodologische begeleiding omdat onbetrouwbare modellen schadelijk kunnen zijn bij het ondersteunen van klinische besluitvorming.

### Hoofdstuk 18. Overzicht van extern gevalideerde voorspelmodellen

Externe validatie is een essentiële stap vóór de klinische toepassing van een voorspelmodel. Deze review vindt 18 studies die 10 van de 59 (17%) ML-modellen valideren. De 18 externe validatiestudies laten onvolledige rapportages van resultaten zien. Dit beperkt een kritische beoordeling van de modelprestaties. Een grotere inspanning is nodig om de beschikbare ML-modellen te testen met externe data in verschillende geografische omgevingen alvorens clinici de modellen volledig kunnen omarmen in de praktijk.



**Figuur 1.** Wereldwijde spreiding van ontwikkeling (n=59) en externe validatie studies (n=18).

### Hoofdstuk 19. Kunstmatige intelligentie versus clinici bij de interpretatie van musculoskeletale beelden

De prestaties van AI-modellen in vergelijking met deze van clinici staan ter discussie. Deze review toont aan dat vergeleken met clinici de AI-modellen, verdeeld over 12 studies, iets beter presteren voor diagnostische nauwkeurigheid en sensitiviteit bij het interpreteren van musculoskeletale afwijkingen. Twee studies evalueren de prestaties van clinici die al dan niet gebruik maken van AI. Beide studies laten aanzienlijke verbeteringen zien in de prestaties van door AI ondersteunde clinici door een afname van 47% van het aantal verkeerde interpretaties en een gemiddelde specificiteit toename van 0.048. AI kan de prestaties van clinici verbeteren als ook gebruik gemaakt wordt van technische aanvulling in plaats van als vervanging.

# AUTHORS AND AFFILIATIONS

**Massachusetts General Hospital, Boston, USA**

*Department of Orthopaedics*

Akhbari, Bardiya
Ashkani-Esfahani, Soheil
Badr, Tim
Bales, John R.
Berner, Emily
Bongers, Michiel E.R.
Clark, Rose
Collins, Austin K.
DiGiovanni, William
Fenn, Brian P.
Ghaednia, Hamid
Harris, Mitchell B.
Horniceck, Francis J.
Hundersmarck, Dennis
Kanbier, Laura N.
Kapoor, Neal D.
Karhade, Aditya V.
Kim, Jason
Kopoulos, Janine
Lans, Amanda
Li, Jacqueline
Lozano-Calderon, Santiago A.
Newman, Erik T.
Oosterhoff, Jacobien H.F.
Ramsey, Duncan C.
Raskin, Kevin A.
Schwab, Joseph H.
Shimizu, Michelle
Shin, David
Sodhi, Alisha
Su, Marie W.
Tomlinson, Emma
Tomson, Timmy
Twining, Peter K.
Yeates, Sarah
Yeung, Caleb M.
Zijlstra, H.
Zhang, Yue

*Department of Radiology*

An, Thomas J.
Bredella, Miriam A.
Buckless, Colleen G.
Rabinov, James D.
Smuclovisky, Eric
Toriani, Martin

*Department of Oncology*

Cohen, Sonia

*Department of Vascular and Endovascular Surgery*

Mohebali, Jahan

*Department of Pathology Laboratory*

Yin, Hung P.

**Harvard Medical School, Boston, USA**

*Department of Epidemiology, Boston*

Beam, Andrew L.

**Brigham and Women's Hospital, Boston, USA**

*Department of Orthopaedics*

Ferrone, Marco L.
Schoenfeld, Andrew J.

**University Medical Center Utrecht, Utrecht, The Netherlands**

*Department of Orthopaedics*

Bindels, Bas J.J.
Ogink, Paul T.
Öner, Fetullah C.

Pielkenrood, Bart J.
Pierik, Robertus J.B.
Tol, Floris R. van
Verlaan, Jorrit-Jan

*Department of Imaging and Cancer*

Verkooijen, Helena M.

*Department of Surgery*

Groot, Vincent P.

*Department of Neurosurgery*

Senders, Joeky T.

**Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands**

*Department of Orthopaedics*

Bramer, Jos A.M.
Dijk, Niek C. van
Janssen, Stein J.
Thio, Quirina C.B.S.

**Erasmus Medical Center, Erasmus University Rotterdam, Rotterdam, the Netherlands**

*Department of Orthopaedics*

Paulino-Pereira, Nuno R.

**University Medical Center Groningen, Groningen, the Netherlands**

*Department of Orthopaedic and Trauma*

Doornberg, Job N.
Groot, T. de

**University of Iowa Hospitals and Clinics, Iowa, USA**

*Department of Orthopaedics*

Gulbrandsen, Trevor R.
Miller, Benjamin J.
Skalitzky, Mary K.

**Memorial Sloan Kettering Cancer Center, New York, USA**

*Department of Neurosurgery*

Bilsky, Mark H.
Laufer, Ilya

**The University of California, Los Angeles, USA**

*Department of Orthopaedics*

Hornicek, Francis J.
Shah, Akash

**The Johns Hopkins University School of Medicine, Baltimore, USA**

*Department of Neurosurgery*

Sciubba, Daniel M.

**The University of Chicago, Illinois, USA**

*Department of Orthopaedics*

El Dafrawy, Mostafa

**National Taiwan University Hospital, Taipei, Taiwan**

*Department of Orthopaedics*

Chen, Chih-Wei
Hu, Ming-Hsiao
Lin, Wei-Hsin
Yang, Jiun-Jen
Yang, Shu-Hua
Yang, Tse-Chuan
Yen, Hung-Kuan

**Seoul National University College of Medicine, Seoul, Korea**

*Department of Orthopaedics*

Han, Ilkyu
Park, Jay H.

**Seoul National University College of Medicine,**
**Seoul, Korea**
*Department of Orthopaedics*
Han, Ilkyu
Park, Jay H.

**Flinders University, Adelaide, Australia**
*Department of Orthopaedics*
Jaarsma, Ruurd L.

# REVIEW COMMITTEE

prof. dr. F.C. Oner (Committee chair)

Professor of Spinal Surgery

University Medical Center Utrecht, The Netherlands

prof. dr. Y. van der Linden

Professor of Palliative Medicine

Leiden University Medical Center, The Netherlands

prof. dr. ir. C.A.T. van den Berg

Professor of Computational Imaging

University Medical Center Utrecht, The Netherlands

prof. dr. S.C.C.M. Teunissen

Professor of Hospice Care

University Medical Center Utrecht, The Netherlands

dr. E.H. Gort

Department of Medical Oncology

University Medical Center Utrecht, The Netherlands

# LIST OF PUBLICATIONS

Hundersmarck D, Groot OQ, Schuijt HJ, Hietbrink F, Leenen LPH, Heng M. Liver Cirrhosis among Patients with Hip Fractures: Liver Dysfunction Dictates Prognosis. *Clin Orthop Relat Res (In press)*

Lans J, Groot OQ, Hazewinkel M, Kaiser P, Lozano-Calderon SA, Heng M, Valerio IL, Eberlin KR. Factors Related to Neuropathic Pain Following Lower Extremity Amputation. *Plast Reconstr Surg. (In press)*

Groot OQ, Lans A, Twining PK, Bongers MER, Kapoor ND, Verlaan JJ, Newman ET, Raskin KA, Lozano-Calderon SA, Janssen SJ, Schwab JH. Clinical Outcome Differences in the Treatment of Impending Versus Completed Pathological Long Bone Fractures. *J Bone Joint Surg Am. (In press)*

Bongers MER, Groot OQ, Buckless CG, Kapoor ND, Twining PK, Schwab JH, Torriani M, Bredella MA. Body Composition Predictors of Mortality in Patients with Spinal Metastases undergoing Surgical Treatment. *Spine J.* Oct 23;S1529-9430

Skalitzky MK, Gulbrandsen TR, Groot OQ, Karhade AV, Verlaan JJ, Schwab JH, Miller BJ. The Preoperative Machine Learning Algorithm for Extremity Metastatic Disease Can Predict 90-day and 1-year Survival: An External Validation Study of 264 Patients. *J of Surg Onc.* Oct 5 (Online ahead of print)

Tseng TE, Lee CC, Yen HK, Groot OQ, Hou CH, Lin SY, Bongers MER, Hu MH, Karhade AV, Ko JC, Lai YH, Yang JJ, Yang RS, Schwab JH, Lin WH. International Validation of the SORG Machine Learning Algorithms for Survival Prediction of Extremity Metastases Undergoing Surgical Treatment. *Clin Orthop Relat Res.* Sept 7 (Online ahead of print)

Ogink PT, Groot OQ, Karhade AV, Bongers MER, Oner FC, Verlaan JJ, Schwab JH. Wide Range of Applications for Machine-Learning Prediction Models in Orthopedic Surgical Outcome: a Systematic Review. *Acta Orthop.* 2021 Oct;92(5):526-531

Paulino Pereira NR, Groot OQ, Verlaan JJ, Bongers MER, Twining PK, Kapoor ND, van Dijk CN, Schwab JH, Bramer JAM. Quality of Life Changes After Surgery for Metastatic Spinal Disease: A Systematic Review and Meta-analysis. *Clin Spine Surg*. 2021 Jun 9 (Online ahead of print)

Groot OQ, Ogink PT, Oosterhoff JHF, Beam AL. Natural Language Processing and Its Role in Spine Surgery: a Narrative review of Potentials and Challenges. *Spine Surgery 2021 Jun 33;2*

Lans A, Oosterhoff JHF, Groot OQ, Fourman MS. Machine Learning Driven Tools in Orthopaedics and Spine Surgery: Hype or Reality? Applications and perception of 31 physician opinions. *Spine Surgery 2021 Jun 33;2*

Oosterhoff JHF, Groot OQ, Thio CBSQ, Bongers MER, Ghaednia H, Karhade AV, Del Fiol G, Kawamoto K. Integration of Automated Dredictive Analytics into Electronic Health Records: Can Spine Surgery Applications Lead the Way using SMART on FHIR and CDS Hooks? *Spine Surgery 2021 Jun 33;2*

Ogink PT, Groot OQ, Bindels JJ, Tobert DG. The use of Machine Learning Prediction Models in Spinal Surgical Outcome: An Overview of Current Development and External Validation Studies. *Spine Surgery 2021 Jun 33;2*

Groot OQ, Bindels JJ, Ogink PT, Kapoor ND, Twining PK, Collins AK, Bongers MER, Lans A, Oosterhoff JHF, Karhade AV, Verlaan JJ, Schwab JH. Availability and Reporting Quality of External Validations of Machine-Learning Prediction Models with Orthopedic Surgical Outcomes: a Systematic Review. *Acta Orthop. 2021 Apr 18;1-9.*

Groot OQ, Paulino Pereira NR, Bongers MER, Ogink PT, Newman ET, Verlaan JJ, Raskin KA, Lozano- Calderon SA, Schwab JH. Do Cohabitants Reliably Complete Questionnaires for Patients in a Terminal Cancer Stage when Assessing Quality of Life, Pain, Depression, and Anxiety? *Clin Orthop Relat Res 2021 Apr 1;479(4);792-801*

Groot OQ, Ogink PT, Lans A, Twining PK, Kapoor ND, DiGiovanni W, Bindels BJJ, Bongers MER, Oosterhoff JHF, Karhade AV, Onfer FC, Verlaan JJ, Schwab JH. Machine Learning Prediction Models in Orthopedic Surgery: a Systematic Ceview in Transparent Reporting. *J Orthop Res 2021 Mar;18*

Groot OQ, Bongers MER, Thio QCBS, Bramer JAM, Verlaan JJ, Newman ET, Raskin KA, Lozano-Calderon SA, Schwab JH. Prospective study for Establishing Minimal Clinically Important Differences in Patients with Surgery for Lower Extremity Metastases. *Acta Oncol 2021 Feb;25;1-7*

Yang JJ, Chen CW, Fourman MS, Bongers MER, Karhade AV, Groot OQ, Lin WH, Yen HK, Huang PH, Yang SH, Schwab JH, Hu MH. International External Validation of the SORG Machine Learning Algorithms for Predicting 90-day and 1-year Survival of Patients with Apine Metastases using a Taiwanese Cohort. *Spine J 2021 Feb; S1529-9430*

Groot OQ, Bongers MER, Karhade AV, Kapoor ND, Fenn BP,Kim J, Verlaan JJ, Schwab JH. Natural Language Processing for Automated Quantification of Bone Metastases Reported in Free-Text Bone Scintigraphy Reports. *Acta Oncol 2020 Dec;59(12);1455-1460*

Groot OQ, Bongers MER, Ogink PT, Senders JT, Bramer JAM, Verlaan JJ, Schwab JH. Does Artificial Intelligence Outperform Natural Intelligence in Interpreting Musculoskeletal Radiological Studies? a Systematic Review. *Clin Orthop Relat Res 2020 Dec;478(12):2751-2764*

Groot OQ, Kapoor ND, Twining PK, Pielkenrood BJ, Bongers MER, Newman ET, Verlaan JJ, Schwab JH. Adipose Tissue Density on CT as a Prognostic Factor in Patients with Cancer: a Systematic Review. *Acta Oncol 2020 Dec;59(12):1488-1495*

Groot OQ, Hundersmarck D, Lans A, Bongers MER, Karhade AV, Zhang Y, van Tol FR, Verlaan JJ, Mohebali J, Schwab JH. Postoperative Adverse Events Secondary to Iatrogenic Vascular Injury during Anterior Lumbar Spinal Surgery. *Spine J 2020 Nov 3;S1529-9430(20)31206-7*

Bongers MER, Karhade AV, Setola E, Gambarotti M, Groot OQ, Erdogan KE, Picci P, Donati DM, Schwab JH, Palmerini E. How Does the Skeletal Oncology Research Group Algorithm's Prediction of 5-year Survival in Patients with Chondrosarcoma Perform on International Validation? *Clin Orthop Relat Res 2020 Oct;478(10):2300-2308*

Oosterhoff JHF, Doornberg JN, Machine Learning Consortium. Artificial Intelligence in Orthopaedics: False Hope or Not? A narrative review along the line of Gartner's hype cycle. *EFORT Open Rev. 2020 Oct; 26; 5(10):593-603*

Karhade AV, Bongers MER, Groot OQ, Kazarian ER, Cha TD, Fogel HA, Hershman SH, Tobert DG, Schoenfeld AJ, Bono CM, Kang JD, Harris MB, Schwab JH. Natural Language Processing for Automated Detection of Incidental Durotomy. *Spine J 2020 May;20(5):695-700*

Bongers MER, Karhade AV, Villavieja J, Groot OQ, Bilsky MH, Laufer I, Schwab JH. Does the SORG Algorithm Generalize to a Contemporary Cohort of Patients with Spinal Metastases on External Validation? *Spine J 2020 May 16;S1529-9430(20)30193-5*

Karhade AV, Bongers MER, Groot OQ, Cha TD, Doorly TP, Fogel HA, Hershman SH, Tobert DG, Srivastav SD, Bono CM, Kang JD, Harris MB, Schwab JH. Development of Machine Learning and Natural Language Processing Algorithms for Preoperative Prediction and Automated Identification of Intraoperative Vascular Injury in Anterior Lumbar Spine Surgery. *Spine J 2020 Apr 12;S1529-9430(20)30135-2*

Karhade AV, Bongers MER, Groot OQ, Cha TD, Doorly TP, Fogel HA, Hershman SH, Tobert DG, Schoenfeld AJ, Kang JD, Harris MB, Bono CM, Schwab JH. Can Natural Language Processing Provide Accurate, Automated Reporting of Wound Infection Requiring Reoperation After Lumbar Discectomy? *Spine J 2020 Mar 4;S1529- 9430(20)30088-7*

Groot OQ, Ogink PT, Paulino Pereira NR, Ferrone ML, Harris MB, Lozano-Calderon SA, Schoenfeld AJ, Schwab JH. High Risk of Symptomatic Venous Thromboembolism After Surgery for Spine Metastatic Bone Lesions: A Retrospective Study. *Clin Orthop Relat Res. 2019;477:1674-1686; "Advances in Motion", Harvard Medical School's newspaper 2019 Aug*

Paulino Pereira NR, Ogink PT, Groot OQ, Ferrone ML, Hornicek FJ, van Dijk CN, Bramer JAM, Schwab JH. Complications and Reoperations after Surgery for 647 Patients with Spine Metastatic Disease. *Spine J. 2019;19:144-156.*

Van Setten J, Warmerdam EG, Groot OQ, de Jonge N, Keating B, Asselbergs FW. Non-HLA Genetic Factors and Their Influence on Heart Transplant Outcomes: a Systematic Review. *Transplant Direct 2019 Jan 21;5(2):e422*

Groot OQ, Ogink PT, Janssen SJ, Paulino Pereira NR, Lozano-Calderon SA, Raskin KA, Hornicek FJ, Schwab JH. High Risk of Venous Thromboembolism After Surgery for Long Bone Metastases: a Retrospective Study of 682 Patients. *Clin Orthop Relat Res. 2018;476:2052-2061.*

The place I will miss most is undoubtedly Walden. The tranquillity, the serenity, the natural surroundings, the dark pond, the dazzling sunset. Thoreau took me there and I hate to leave, each and every time I visit for a swim in the cool water, for a quiet and leisurely afternoon break during my PhD..

Appleton 88 / Chandler Street / 02116 / Yawk
Lubberts / Phoenix Arizona USTA 4.5 Natio
Thor / Goatee / Las Vegas peukie roken / F
Jigglypuff / Harvard MD PhD / Scotty kluiz
Berlin / Pistol Pete / Honduran handshake /
You can't see me / Over de schutting bij Bono gooien / The Donald / Build that wall / Clinton *** / Lock her
rocket man / Prof. Bongers / Baking Banging Bongers / Stennis Pechvogelhuis / De wissel / Timmy Tomson
Ben niet verslaafd, maar heb het wel nodig / Een druppeltje / Walden Pond / Safety first kids / Haute Coffee/
winnaars / NLP'tje / The Office / Yeppers / Liberty Hotel / JFK Library / Cheese, office, and puzzle maratho
rave cave / Hadrianus / Tivoli / When they go low, we go high / Mango aapje / Peukje roken / Monnikenkap
Container city / Rise 'n shine with lettuce and tomato / Freedom trail / Minute man trail / Brian of Tarth / Cer
of Ljubljana / Lake Bled paintings / Cornelis Krusemanstraat 18-3 / Hoendiepstraat 36-2 / TB12 / Wateerrrr /
Colombier / Hymer Hummer / Paradise on wheels / Second love / No hablo espanol / Fuego fuego / Olivier s
Bellevue / 11:28 Lac d'Aiguebelette / Reinier Zonneveld / Camping Bagatelle le Pavillon Bleu / Camping el Llac
/ King Leonidas / Schaken / Nadine's Gambit / Tinto de verrano / Divorce / Freak show DK / Yeabridge green
de France - Château Fontainebleau - Nitry - Noyers- Rathier - Château Chalon - Baume-les-Messieurs - Lausa
- Lac d'Aiguebelette - Avignon - Pont du Gard - Châteuaneuf-du-Pape - Saint-Rémy-de-Provence - Les Beau
- Granada - Malaga - Ronda - Seville - Cadiz - Montbron - Village Le Chat - Den Haag / Baja California / To
/ Ipo / Wellington / Villa Bugambilias / Cabo / Unhinged left wing mob / Sleepy sleezy Joe / Napoleon the Gr
Puck / Muilkorf / The roast: a tribute to Mark / Miss Grande / Carrot juice / Lobsters rolls / Amber & Dann
Denali Dome Home b&b / Coach / Marmot / Mt. Healy / Denali National Park / Moose in parking lot / Sewa
lodge / Kariboe / Black bear / Donald the pastor / Savage alpine trail / Mt. Denali / French toast / Mt. marath
/ Captain Bob / Ocean is the motion / Jen / Sue and Glenn +5z's / Wolverine / Bernard Explorer / 5 gallon w
$600 / Daniel / Beluga whales / House of cards / Sea of Cortez / Zodiac / HMS *hole / Alaska dad body / Ame
Eagles/ Sluice Juice / Josh, Mikey Keane / Rod Laver Cup / Coastie Rich / GH885 / CS521 / CS677 / Avi / MI
NOT a gun free zone / Foliage / Tuben / He HK / Study godfather / Study sugar daddy / Unbelievable / To be

ey / Deluxe / Angel / Tuna melt / Steve / Damian / Steve Power / Stay healthy bruv / Brady Buccaneer / Coach
onal Tournament / Mt. Lemmon / Tucson / Big, beautiful wall / Mr. And Miss Lane / Chip & Cherry / Captain
Paultje Pogink / Cobus/ It nibbles the bone away / Deluxe / Tasty / Xiang Pao / Brier score / Ben & Jerry's /
zenaar / Thuiswerkdagje / Cape Cod / Crackhead Brody / Utila / Dr. John / Ome John / Sniffer / Aconchego /
/ Paultje Prepaid / Chef van Tol / Centibindels / Huizen ruikende rikkel / Patriots - Falcons 28-3 / Silentium /
up / USA / Four more years / crazy Bernie / Total snoozefest / low IQ democrat / 12 angry democrats / Little
n / Miss Yeates / Master of Coin / Burrito bar / Jambalaya / Cho Lei Mein / Kaulo mooi / IJskoude corona /
Hotline K. / Even Irene bellen / Chinese virus / Huiliang / Young Jeezy / Ankle and Foot / Iraanse Nobelprijs
n / Ja, exact / Yes, exactly / Mitchie / Koffiezetten kijken / Vogeltjes kijken / Weight gainer Stennis / Libyan
sel / Hel Box / Beaurreaux / Bud light with lime / Chipotleaway / Someone took a nap in the shower again /
cle / Lizard man / Boris Brejcha / HJH / Fakear / El Toro loco / Drieluik / COVID-19 / Alternative facts / Heart
That's what she said / De wereld volgens Gijp / Stokbrood tand / Sacrofaag / Exploding kittens / Le Grand
upertramp / Into the Wild / De Bourgondiërs / TCS camping Genève-Vesenaz / Les Rives du Lac / Camping
/ Camping Coll Vert / Camping La Volta / Ole / Las Lomas / El Sur / Playa las Duna / San Anton / Miss Kitty
287 / Farrow & Ball / 1993 GN-TS-60 / Den Haag - Amsterdam - Tessenderlo - Tongeren - Champagne - Ile
anne - Lake Genève - Genève - Yvoire - Nyon - Lac Annecy - Col du Semnoz - Col du Leschaux - Bellecombe
ux-de-Provence - Arles - Abbey Frigolet - Lago Banolas - Barcelona - Castell de Peníscola - Valencia - Oliva
dos Santos / Matt / Genata / Emilie / Grey whales / Swinging from d to d / Frankie / Harold / Georgina / Bob
reat / de Bourgondiërs / Sheela / Criky it is a feisty one / Run, runnn / King Henry VII / Doctor Jay / Rambo
y / Emma's keys / The Cheesemonger / Rhonda / Pat / Cult leader / Historic anchorage hotel / Anchorage/
rd / Bobby / Girdlewood / 49th State brewery / Magic bus / The Grey / Into the wild / Snatch / Midnight sun
non / Pitt bar / Psycho Pete / Nome / Board of trade / Moonlight water well / Tod / Derrek / Rolland / David
ater trips / Musk ox / Frankenstein dredges / Sledge island / Font street / Hooper Bay / Jimmy / Walrus tusk
lia / Narni / First class / AAOS San Diego / Santa Barbara / Jeff / Carla / Nelly / Sequence / Code names / BC
T courts / Boston Common courts / Carter Playground / Murph Challenge / Bethlehem / Sandwich / this is
e honest / Hypnotized / Laat ik het zo zeggen / Halloween pub crawl / Curacao...

### De Animalium Proprietate

Social media makes one forget the charm of writing letters. While studying abroad, first in Portland, later in Boston, I rediscovered the alluring taste of sending postcards to family and friends. Over the years, my collection of cards holds hundreds of sweet memories of artifacts, monuments, historical sites, natural marvels and cultural highlights. Each card and its message present a personal sentiment of a cherished experience, big and small.

One of my favorite illustrations will always be a weirdly goggling creature in a 16th century manuscript seen at the Isabella Stewart Gardner Museum in Boston. At the 2016 exhibition 'Beyond Words', the image of the hammerhead shark evokes a world of beastly wonders. It captures a creative and vivid effort to explore and to educate just as Manuel Philes intended. This Byzantine poet wrote a poem on the characteristics of animals *De animalium proprietate* around 1320. More than 200 years later, the poem was illuminated in an Italian Renaissance book by a magnificent pen and ink drawing with color wash. The manuscript is to be seen in the Houghton Library, Harvard University.

The same intriguing picture now illuminates my publication on bone metastases. We no longer marvel at sea creatures, but the unknown is still unnerving. That is also the case for artificial intelligence being a new and promising field. Together with my colleagues in Boston and Utrecht we explored the possibilities of artificial intelligence to help improving treatment of patients with bone metastases. We also recognized the caveats related to transparency of data and/or code, clinical usefulness, bias, and socioeconomic inequality. However, we trust artificial intelligence will develop into a useful and trustworthy tool in the very near future. I would be proud to contribute to that purpose to the benefit of the patients.

# Propositions

Belonging to the thesis

## Improving Patient Selection for Surgical Treatment of Bone Metastases

The number of patients with bone metastases undergoing surgery is increasing, inviting the detrimental outcomes of index hospitalization and postoperative adverse events.
*This thesis*

Multidisciplinary approaches are needed to formulate individualized plans while also contemplating non-operative strategies in these patients at high-risk of poor outcomes.
*This thesis*

An easy to use, accurate, and validated prediction tool which identifies if a patient with a metastatic bone lesion is at risk for developing a pathological fracture would be valuable.
*This thesis*

Artificial intelligence algorithms are rapidly emerging tools in medicine and a potential resource to improve patient selection and prognostication for surgical intervention.
*This thesis*

Of the existing machine learning prediction models in orthopaedic surgery, over 40% are at high risk of bias and only 17% are externally validated.
*This thesis*

Surgeons, together with patients, must weigh the likelihood of both benefits and adverse events when choosing a surgical intervention – artificial intelligence can aid this decision-making process.
*This thesis*

Clinicians should embrace artificial intelligence as a supplement by creating a desirable synergy between "human and machine" rather than the ancient opposition of "human versus machine."
*This thesis*

# UITNODIGING

U bent van harte uitgenodigd voor
het bijwonen van de openbare
verdediging van het proefschrift

## Improving Patient Selection for Surgical Treatment of Bone Metastases

door Olivier Q. Groot

Donderdag 23 december 2021 om
16.15u in het Academiegebouw,
Domplein 29 in Utrecht

### PARANIMFEN

Michiel E.R. Bongers

Paulus T. Ogink

### Olivier Q. Groot

Hoendiepstraat 36-2
1079LS Amsterdam
oqgroot@gmail.com

U wordt verzocht 15 minuten van
tevoren aanwezig te zijn. Deuren
sluiten om precies 16.15u.

# INVITATION

You are kindly invited to attend the public defence of the PhD thesis

## Improving Patient Selection for Surgical Treatment of Bone Metastases

by Olivier Q. Groot

Thursday, December 23rd 2021 at 4.15pm in the University Hall (Academiegebouw), Domplein 29 in Utrecht

### PARANIMPHS

Michiel E.R. Bongers

Paulus T. Ogink

Olivier Q. Groot
Hoendiepstraat 36-2
1079LS Amsterdam
oqgroot@gmail.com

You are kindly requested to be present 15 minutes prior to the start of the ceremony. Doors will close at precisely 4.15pm.