From survival prediction to treatment decision
in lung cancer

Wouter A.C. van Amsterdam

From survival prediction to treatment decision in lung cancer

Van het voorspellen van overleving naar behandelbeslissingen in longkanker
(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 22 september 2022 des middags te 12.15 uur

door

Wouter Anton Christiaan van Amsterdam

geboren op 16 april 1989
te Doetinchem

Financial support by “National Fonds Tegen Kanker” for the publication of this thesis is gratefully acknowledged.

Lay-out: Wouter A.C. van Amsterdam & Dennis Hendriks (ProefschriftMaken.nl)
Printed: ProefschriftMaken.nl
Cover design: Wouter A.C. van Amsterdam
Cover image: Generated using DALL-E 2 and Stable Diffusion, with variations of the prompt:
“the road bifurcates and branches left and right at a crossroads around a tree, view from
behind of an old person standing the crossroads, a large tree in the shape of lungs stands in
front of him. Scientific American cover with colours of Zion national park, trending on
Artstation HQ”

ISBN: 978 94 642 3888 4

© Wouter van Amsterdam, 2022
All rights reserved. No part of this thesis may be reproduced, stored in a retrieval system, or transmitted in any other form or by any other means (e.g. mechanically, by photocopy, by recording, or otherwise), without permission from the author.

From survival prediction to treatment decision in lung cancer

Wouter Anton Christiaan van Amsterdam

Promotoren:
Prof. dr. P.A. de Jong
Prof. dr. T. Leiner

Copromotoren:
Dr. R. Ranganath
Dr. J.J.C. Verhoeff

Beoordelingscommissie:
Dr. J.A. De Langen
Prof. dr. D.L. Oberski
Prof. dr. R.H.H. Groenwold
Prof. dr. P.J. van Diest
Prof. dr. M. Koopman (voorzitter)

Content list

Chapter 1: Introduction 7

Part I: **Predicting overall survival for non-small cell lung cancer patients** **13**

Chapter 2: Prognostic factors for overall survival of stage III non-small cell lung cancer patients on computed tomography: A systematic review and meta-analysis 15
Radiotherapy and Oncology, 2020

Chapter 3: The association between psoas volume and overall survival depends on psoas radiodensity: a cohort study in non-small cell lung cancer 47
Journal of Personalized Medicine, 2022

Part II: **Improving treatment decisions with individual treatment effect estimates** **65**

Chapter 4: Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning 67
npj Digital Medicine 2019

Chapter 5: Individual treatment effect estimation in the presence of unobserved confounding using proxies: a cohort study in stage III non-small cell lung cancer 77
Scientific Reports 2022

Chapter 6: Conditional average treatment effect estimation with treatment offset models 117
Submitted

Chapter 7: Decision making in cancer: causal questions require causal answers 135
Submitted

Chapter 8: Discussion 145

Chapter 9: General summary 159

Chapter 10: Nederlandse Samenvatting 165

List of publications 172
Dankwoord 173
Curriculum Vitae 175

CHAPTER 1

Introduction

Lung cancer is the greatest cause of cancer related death, both worldwide (1) and in the Netherlands (2). This is because lung cancer has a high incidence, meaning that it frequently occurs, and has a high mortality, meaning that it often leads to death. In half of the lung cancer patients in the Netherlands, the disease has already spread to other organs at the time of diagnosis (3). Curative treatment is no longer an option for these patients. Depending on how far the disease has spread, lung cancer is divided in four stages using the Tumor Node and Metastasis classification system (4) defined by the American Joint Committee on Cancer. Disease stage at the time of diagnosis is important for the available treatment options and for overall survival, with 5-year overall survival rates of 53% for stage I, 38% for stage II, 17% for stage III and 3% for stage IV (2). However, these are group-level statistics and there are large differences in survival between patients, even within a single stage. Once the diagnosis and stage are known for a specific patient two crucial questions remain: what is the prognosis? and what treatment options are available that have the best chance of improving the prognosis? Ideally, these two questions are answered on the level of the individual patient, not at the group level.

Individual prognosis: no two tumors are the same, no two patients are the same

There is great variability in overall survival between patients even within a single disease stage. For instance, for stage II lung cancer, 25% of patients will die within 1 year of diagnosis whereas another 25% of patients will still be alive after 10 years (5). Part of the difference in survival will be due to pure chance. Indeed, random mutations cause cancer in the first place and play an important role cancer progression (6,7). Furthermore, overall survival depends on chance-events such as contracting infectious diseases, accidents and other unforeseeable events. Despite this intrinsic randomness, part of the variation in overall survival will be explainable based on characteristics of the tumor and the patient that can be measured at the time of the diagnosis. There is much research interest in discovering new characteristics of tumors that are related to patient prognosis. These tumor characteristics can be based on different types of medical examination, including physical examination, histologic examination, radiologic imaging, genetic profiling and molecular profiling. In addition to variation in the characteristics of the tumor, there are important differences between patients as well. If there are two patients with the exact same lung tumor but one patient is younger and in better overall health than the other, the younger patient will survive longer on average. A large field of research is dedicated to predicting patient outcomes using tumor and patient characteristics. Prognosis predictions are useful for patient counseling.

Individual treatment: all treatments are not equally effective for all patients and tumors

Prognosis predictions are relevant for patient counseling but do not answer what is arguably the most pressing question in cancer care: what treatment will have the best positive impact on the health outcome of an individual patient. The gold standard for estimating the effects of treatments are randomized controlled trials. In randomized controlled trials, patients are randomly assigned to receive one of the possible treatments under study. In a well conducted randomized trial, due to the randomization, any differences in health outcomes between the treatment groups is attributable to the causal effect of the treatment. A downside of randomized trials is that they generally estimate the treatment effect at the group level, whereas it is likely that different treatments are not equally effective for all patient subgroups. Whereas randomized trials estimate the *average treatment effect,* patients and doctors are more interested in the *individual treatment effect*: "what is the effectiveness of this treatment, given that we know the characteristics of this patient?". Randomized controlled trials often do not include sufficient patients to estimate the individual treatment effect. Observational studies, where patients are treated according to the standard clinical care, generally include more patients than randomized trials and can potentially provide evidence on the effectiveness of treatments in different subgroups. As opposed to randomized trials, patients in observational studies who receive different treatments often differ with respect to important characteristics that also influence their outcomes, so called 'confounders'. As a concrete example, stage I lung cancer patients are often treated with surgery, unless they are unfit for surgery in which case they are treated with radiotherapy. The older and weaker patients who are treated with radiotherapy have much worse overall survival than the patients treated surgery, but this difference in survival is not attributable to the causal effect of surgery versus radiotherapy, as the patient groups were not comparable to begin with. However, if two patients who receive different treatments are similar with respect to all confounders, the difference in outcome between these patients is attributable to the causal treatment effect. As a result, estimating treatment effects from observational data has an important complicating factor: the presence of confounders. Given the importance of estimating individual treatment effects, and given that randomized trials often provide insufficient evidence for this, methods that allow for inferring treatment effects from observational studies are of crucial importance.

Outline of this thesis

This thesis is composed of two parts: in **part 1** we present two studies on predicting prognosis for non-small cell lung cancer patients. The first is a summary of published literature that studies tumor characteristics that are visible on computed tomography scans of lung cancer patients. The second study focusses on patient characteristics and presents a new hypothesis on how the amount of muscle tissue and the density of the muscle tissue of a patient may be related to overall survival. In **part 2** we present three studies that go one step further and estimate what the best treatment option is for an individual patient given their characteristics. The first study uses an advanced statistical method called 'deep learning' to estimate the prognosis of a patient and the treatment effect based on medical images in the presence of a challenging causal problem called 'collider bias'. In the second study we address an important issue in observational cancer research: the presence of unobserved confounders. This study presents a new method to estimate treatment effects when there are unobserved confounders but there are proxy measurements available. The third study investigates a method to convert treatment effect estimates on a relative scale to conditional average treatment effect estimates using the baseline risk of a patient. We finish with a comment that emphasizes that supporting treatment decisions is a causal task and thus requires causal approaches. Whereas much outcome prediction research is motivated by supporting future treatment decisions, the causal dimension of this task is often ignored, leading to substantial risk of harm. We elaborate on what is needed to conduct prediction research that is useful for supporting future decisions.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
2. NKR Cijfers [Internet]. [cited 2021 Aug 25]. Available from: https://iknl.nl/nkr-cijfers
3. Helft van longkankerpatiënten heeft gevorderde kanker bij diagnose [Internet]. [cited 2021 Aug 25]. Available from: https://iknl.nl/nieuws/2018/helft-van-longkankerpatienten-heeft-gevorderde-kan
4. TNM Classification of Malignant Tumours, 7th Edition [Internet]. Wiley.com. [cited 2020 Dec 1]. Available from: https://www.wiley.com/en-nl/TNM+Classification+of+Malignant+Tumours%2C+7th+Edition-p-9781444358964
5. Overleving longkanker [Internet]. [cited 2021 Sep 10]. Available from: https://iknl.nl/kankersoorten/longkanker/registratie/overleving
6. Tomasetti C, Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science. 2015 Jan 2;347(6217):78–81.
7. Lucio L, Paolo PP. Causality and Chance in the Development of Cancer. N Engl J Med. 2015;5.

PART I

Predicting overall survival for non-small cell lung cancer patients

CHAPTER 2

Prognostic factors for overall survival of stage III non-small cell lung cancer patients on computed tomography: A systematic review and meta-analysis

Myra van Laar, Wouter A.C. van Amsterdam, Anne S.R. van Lindert, Pim A. de Jong, Joost J.C. Verhoeff

Radiotherapy and Oncology 151 (2020) 152–175
https://doi.org/10.1016/j.radonc.2020.07.030

Abstract

Introduction: Prognosis prediction is central in treatment decision making and quality of life for nonsmall cell lung cancer (NSCLC) patients. However, conventional computed tomography (CT) related prognostic factors may not apply to the challenging stage III NSCLC group. The aim of this systematic review was therefore to identify and evaluate CT-related prognostic factors for overall survival (OS) of stage III NSCLC.

Methods: The Medline, Embase, and Cochrane electronic databases were searched. After study selection, risk of bias was estimated for the included studies. Meta-analysis of univariate results was performed when sufficient data were available.

Results: 1595 of the 11,996 retrieved records were selected for full text review, leading to inclusion of 65 studies that reported data of 144,513 stage III NSCLC patients andcompromising 26 unique CT-related prognostic factors. Relevance and validity varied substantially, few studies had low relevance and validity. Only four studies evaluated the added value of new prognostic factors compared with recognized clinical factors. Included studies suggested gross tumor volume (meta-analysis: HR = 1.22, 95%CI: 1.05–1.42), tumor diameter, nodal volume, and pleural effusion, are prognostic in patients treated with chemoradiation. Clinical T-stage and location (right/left) were likely not prognostic within stage III NSCLC. Inconclusive are several radiomic features, tumor volume, atelectasis, location (pulmonary lobes, central/peripheral), interstitial lung abnormalities, great vessel invasion, pit-fall sign, and cavitation.

Conclusions: Tumor-size and nodal size-related factors are prognostic for OS in stage III NSCLC. Future studies should carefully report study characteristics and contrast factors with guideline recognized factors to improve evidence evaluation and validation.

Cancer is a major cause of mortality and a societal burden, which poses a medical challenge to this day [1]. Lung cancer is one of the most common types of cancer with respect to incidence [2]. The relatively low survival of lung cancer in conjunction with treatment induced toxicity emphasizes the importance of considering prognosis before making treatment decisions [1–7]. Stage III non-small cell lung cancer (NSCLC) compromises a particularly difficult subgroup in this regard, because it represents a heterogeneous group of patients. Trials conducted in the last decade show improved survival outcomes compared to older trials, resulting from introduction of PET-CT and MRI for optimal staging ('stage migration') and from improvements in surgical treatment, radiotherapy, and introduction of immunotherapy. Still, only a proportion of all patients benefit from these intensive multimodality treatment schemes and a significant proportion experiences toxicity. This is the challenge presented to multidisciplinary boards: balancing the chance of disease curation and quality of life, making treatment decisions while taking into account risk factors as individual prognostic factors. Current guidelines acknowledge several prognostic factors including stage at diagnosis, performance status, gender, and weight loss [8]. Prognostic factors can also be derived by medical imaging modalities. Of all modalities used in diagnosis and staging of NSCLC, computed tomography (CT) is most commonly used [9]. CT, typically used to obtain information on tumor size and location, is integral for determination of clinical T-stage and N-stage [9,10]. In recent years an abundance of articles considering factors for overall survival (OS) that can be measured by CT has been published [11–21]. In order for these CT-related prognostic factors to become applicable in clinical practice, a clear overview should be created. Other common outcomes are progression-free and disease-free survival. We note these outcomes are mainly of interest for comparing treatment efficacy. OS is arguably the most relevant outcome from a patient perspective, therefore this review focusses specifically on OS. For these reasons, the aim of this study was to systematically review and appraise the evidence on CT-related prognostic factors for OS of stage III NSCLC patients, and to synthesize the evidence with a meta-analysis where possible.

Methods

Search strategy

This study was pre-registered in the PROSPERO registration of systematic reviews (registration number/ID 160936). The Medline (via PubMed), Embase, and Cochrane electronic databases were searched for literature (last queried on 30-09-2019). The search terms consisted of terms reflecting domain, determinant, and outcome of the research question. The complete queries are available in Appendix B.

Study selection

Studies retrieved by this search term were screened on title and abstract using the online screening tool Abstrackr [22]. A blinded pilot title/abstract screen of 100 articles was completed by 2 independent reviewers (MvL, WA), conflicts were resolved via consensus by the 2 reviewers. During the following full text review selected publications, reviews, and editorials were screened for cross-references. Original studies discussing the effect of a prognostic factor for OS that can be measured on CT prior to treatment allocation of stage III NSCLC patients were included. Excluded were studies not including stage III NSCLC patients, considering no CTrelated prognostic factors for stage III NSCLC patients, written in a language other than English, French, German, or Dutch, and studies that explicitly stated consisting of only pathological staged patients (n = 17), because initial treatment decisions can only be based on clinical stage [9]. The utilized TNM-staging system was used as a relevance criterion. Additionally, when multiple studies explicitly stated use of the same patient cohort, the publication with the most recent data was included. Finally, results of multivariable analyses of studies containing a stage III patient number per variable below 5 or containing variables measured after treatment initiation (n = 6), were excluded from analysis.

Data collection

Data was extracted from the inclusions with a data extraction sheet based on the Cochrane Handbook [23], which was piloted for 2 randomly selected publications. After some adjustments were made during a consensus meeting (MvL, WA, JV), the final version (Appendix D) was used to extract data regarding baseline characteristics, treatment, prognostic factors, outcome measures, and general study information, including the utilized TNM-staging system edition. Where the utilized edition was not specified, an estimate was made based on both inclusion period and references of the article. In cases where outcome measures were reported as model coefficient, the hazard ratio was calculated by exponentiating the model coefficient.

Risk of bias assessment

A grading system for critical appraisal was designed, based on the SIGN and TRIPOD [24,25], to separately assess relevance and validity of included publications on outcome level. After piloting for 4 random inclusions, some adjustments were made during a consensus meeting (MvL, WA, JV), giving rise to the final version (Appendix C), which was used to assess both relevance and validity of all inclusions.

Statistics

A meta-analysis was performed on model coefficients from univariate models of prognostic factors when three or more studies reported at least either: the HR and associated standard error, HR and p-value, or a confidence interval. When the reported HR was numerically identical to either the upper bound or lower bound of the confidence interval (e.g. due to rounding), this study was excluded from the meta-analysis.

As the Cox proportional hazards model models the hazards as log-hazard ratios, we log-transformed all hazard ratios, standard errors, and confidence intervals before pooling. When not directly reported, the standard error of the log-HR was recalculated using the range of the log-HR confidence interval (upper minus lower) divided by 3.92 (which is the number of standard deviations included in the 95% confidence interval). If the absolute difference between the upper and the lower was less than 0.05 (leading to numerical inaccuracies due to rounding), or when the CI was not reported but the *p*-value was, we recalculated the standard error using the *p*-value. For this calculation we assumed that the pvalue was calculated based on a Chi-square distribution with one degree of freedom on the Wald-statistic, which is the default method for calculating the p-value in most statistical software packages. For continuous prognostic factors, the log-HR was standardized to a similar unit of measurement. As the included studies ranged a wide period of inclusion times, different TNM staging methods, and different treatment modalities, we used a randomeffects model to pool results, utilizing the Paule-Mandel method for estimating between study variance τ [26]. In addition, between study heterogeneity was estimated using Higgin's & Tompson's I^2 [27]. We did not perform meta-regression, nor did we perform the Egger's test for publication bias as the number of studies was <10 for each comparison [28]. The meta-analysis was performed in R, version 3.6.3, using packages 'meta' and 'dmetar' [29,30].

Results

A total of 11,996 records were retrieved (519 duplicates; Fig. 1), consisting of 10,108 results on Medline, 1863 on Embase, and 25 on Cochrane. The 1595 publications selected for full text review yielded 53 original publications, 8 reviews and 3 editorials. After searching cross-references, a total of 65 original publications were included.

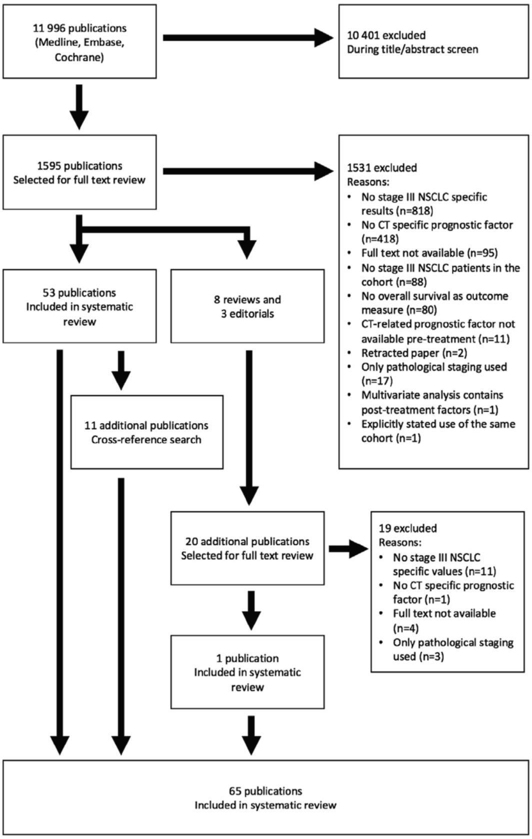

11 996 publications
(Medline, Embase, Cochrane)
10 401 excluded
During title/abstract screen
1595 publications
Selected for full text review
1531 excluded
Reasons:
• No stage III NSCLC specific results (n=818)
• No CT specific prognostic factor (n=418)
• Full text not available (n=95)
• No stage III NSCLC patients in the cohort (n=88)
• No overall survival as outcome measure (n=80)
• CT-related prognostic factor not available pre-treatment (n=11)
• Retracted paper (n=2)
• Only pathological staging used (n=17)
• Multivariate analysis contains post-treatment factors (n=1)
• Explicitly stated use of the same cohort (n=1)
53 publications
Included in systematic review
8 reviews and
3 editorials
11 additional publications
Cross-reference search
20 additional publications
Selected for full text review
19 excluded
Reasons:
• No stage III NSCLC specific values (n=11)
• No CT specific prognostic factor (n=1)
• Full text not available (n=4)
• Only pathological staging used (n=3)
1 publication
Included in systematic review
65 publications
Included in systematic review

Fig. 1. Flow chart of study selection: Flow chart of study selection from the Medline, Embase, and Cochrane database.
Abbreviations: CT: Computed tomography, NSCLC: Non-small cell lung cancer.

The 65 inclusions reported data of 144,513 stage III NSCLC patients (112,082 reported stage IIIA, 31,888 IIIB, and 53 IIIC; Table 1). These studies yielded a total of 26 unique CT-related prognostic factors. Most studies had a retrospective cohort study design; nine studies reported a prospective cohort study design [14,20,31– 37]. In studies reporting follow-up duration, median follow-up ranged from 10 to 70.8 months. Thirty inclusions explicitly stated using a clinical staging method [12,14,16,21,31,33,36,38–60], while a combination of clinical and pathological staging was used in 3 studies [61–63]. The remaining 32 studies did not specify the staging method [11,13,15,17,18,20,32,34,35,37,64–85]. More recent staging systems TNM6 (2002, n = 9), TNM7 (2009, n = 23), and TNM8 (2017, n = 3) were used in 35 publications [11–13,15,16,33,35,36,38,39,42,46,47,49,51,52,54–57,59,61–63,65,66,68,72,73,75,76,78,80,81,85]. Use of less recent staging systems, such as TNM4 [48,67] and TNM5 [14,43,74,79], was stated in 6 inclusions. In critical appraisal, studies that made use of the less recent staging systems were considered to be less relevant. The remaining 24 publications did not explicitly report the utilized staging system.

The stage III cohort generally consisted of multiple histological types with a majority of squamous cell carcinoma and adenocarcinoma patients, except for 2 studies which consisted solely of adenocarcinoma [17] or squamous cell carcinoma[51] patients and 10 studies in which histological type was not reported specifically for stage III patients [14,31,33,35,39,74] or at all [20,50,64,70]. Stage III patients were treated exclusively with chemotherapy and/or radiotherapy in 39 studies [11–13,15,18,20,21,31,32,34–36,41,45,47,48,51,53–56,58–60,64,65,67,70–72,74,76,77,80–85], while surgery was an option in 25 publications [14,16,17,33,37–40,42–44,46,49, 50,52,57,61–63,66,68,73,75,78,79]. A single study did not report treatment modalities [69].

Finally, it should be taken into account that 4 studies made use of the Surveillance, Epidemiology, and End Results (SEER) database with a similar inclusion period and studied the same prognostic factor, indicating that their data is likely to overlap [49,68,69,78]. Overlap of recruitment period and measured prognostic factors was also present in 5 cohort studies that took place at the MD Anderson Cancer Center [11,35,41,65,76], and 2 at the Stanford University School of Medicine [64,70] and National Cancer Center Hospital East [45,52]. The results of studies with presumed overlapping data, taking the individual relevance and validity of the studies into consideration, were treated as results of a single study in data analysis.

The score for relevance ranged from low to high. Five publications were considered to have a high and 8 a low relevance, 52 a medium (Appendix Figure C.3, Table C.3). Low relevance was assigned due to a lack of explicit description of patient characteristics in the stage III cohort [32,50,69], or pronounced discrepancies with the standard stage III population [17,61–63,85]. Most studies (*n* = 62) were estimated to have a medium validity. Two studies were assessed to have a high validity [13,71] and 1 study had an estimated low validity, as it did not report confidence interval (CI) or p-values [11]. In Appendix C results of critical appraisal are described in more detail.

In the 65 inclusions, 26 individual CT-related prognostic factors for OS of stage III NSCLC patients were described. These 26 factors were divided in 5 categories: Radiomic features (Homogeneity, Kurtosis, Standard deviation, Entropy, Skewness, Mean HU, Largest axial slice average, Average, Largest axial slice uniformity, Busyness, Infomc1, Sosvariance), Size-related prognostic factors excluding T-stage (Tumor diameter, Tumor volume, Gross Tumor Volume), T-stage, Nodal factors (Lymph node volume, Lymph node diameter), and Other CT-related prognostic factors (Atelectasis, Location, Cavitation, Cavitary wall thickness, Interstitial lung abnormalities, Great vessel invasion, Pit-fall sign, Pleural effusion).

Two inclusions studied radiomic features, yielding 12 individual prognostic factors (Table 2) [11,12]. Both studies consisted of stage IIIA and IIIB patients treated with concurrent chemoradiation with a similar distribution of histological subtypes to other inclusions. The association between homogeneity, kurtosis, standard deviation, entropy, skewness, and mean Hounsfield unit (HU) and OS was studied in a single publication. Entropy and skewness were calculated from the HU-histogram. Entropy reflects irregularity in HU-values, while skewness reflects asymmetry of the histogram. While homogeneity, kurtosis, and standard deviation were not significant on univariate analysis, entropy, skewness, and mean HU were significant in both univariate and multivariable analysis [12]. In the second study, 8 radiomic features were measured on either contrast enhanced or 4D-CT scans giving rise to average intensity projection and expiratory phase images. Considering the diverse measurement techniques (LoG, IHIST, GRAD, NGTDM, COM), outcomes of 12 unique factors were reported as coefficients in a model for OS. This model was reported to be significantly better than the model containing solely conventional prognostic factors. However, metrics regarding individual statistical significance were not reported, meaning that while included factors are likely to be significant, their individual prognostic value remains uncertain. It should also be noted that the first study did not specify the utilized software, meaning that comparability of the 2 studies is decreased [11]. In summary, both inclusions indicated radiomic features with potential, including entropy, skewness, mean HU, largest axial slice average, largest axial slice uniformity, HU kurtosis, HU infomc1, HU standard deviation, and HU sosvariance, which should be validated in larger cohorts.

Three size-related prognostic factors were found across 38 inclusions (Table 3): Tumor diameter [16,32,34,38,45,48–52,57,58,60,68,69,78,83], Tumor volume [16,18,66,67,71,73,80,81], and Gross Tumor Volume (GTV) [11,13,15,18,21,35,41,47,54,64,65,70, 73,74,76,77,80]. Tumor volume and GTV were considered unique prognostic factors, as GTV encompasses both the volume of the primary tumor and involved lymph nodes [11,13,15,18,41,64,65], where tumor volume includes only primary tumor volume [13,16,66,67,71].

Table 1. Study characteristics.

7 Study citation (PMID)	Study duration (start – end date)	Country	Source of data	Study design	CT-related prognostic factor	Number of stage III participants	TNM staging system used (estimation)	Treatment of the stage III NSCLC participants	NSCLC histological subtypes present	Outcome measure (follow-up)
Huo X, 2017 (29441096)	2005–2011	China	Cohort (The Second Hospital of Tianjin Medical University)	Retrospective	T-stage, Tumor diameter	Clinical stage: IIIA (26), IIIB (156)	NR (TNM6 or TNM7)	Iodine-125 seed implantation and chemotherapy	Squamous cell carcinoma (113), Adenocarcinoma (62), Not specified (7)	Median follow-up: 23 months
Li M, 2004 (15541820)	1994–1998	Japan	Cohort (Hospital of Jiangsu University)	Prospective	Pit-fall sign	Clinical stage: IIIA (10), IIIB (6)	TNM5	Surgery	Not specified for stage III; Overall study population: Squamous cell carcinoma (11), Adenocarcinoma (90), Large cell carcinoma (1), Adenosquamous cell carcinoma (1)	NR
Yilmaz U, 2018 (29559214)	2008–2015	Turkey	Cohort (Dr Suat Seren Chest Disease and Surgery Training and Research Hospital)	Retrospective	T-stage	Clinical stage: IIIA (20), IIIB (49), IIIC (10)	TNM8	Concurrent chemoradiotherapy	Squamous cell carcinoma (58), Not specified (21)	Median follow-up: 20.7 months
Firat S, 2002 (12243808)	1983–1991	USA	Cohort (4 RTOGstudies)	Retrospective	Tumor diameter	Clinical stage: IIIA (69), IIIB (43)	NR (TNM2 or TNM3)	Radiotherapy	Squamous cell carcinoma (79), Not specified (33)	NR
Lee HY, 2012 (22265854)	2004–2009	South Korea	Cohort (Samsung Medical Center)	Retrospective	Tumor diameter	Clinical stage: IIIAN2 (205)	TNM7	Neoadjuvant chemoradiotherapy and surgery	Squamous cell carcinoma (82), Adenocarcinoma (112), Large cell/ neuroendocrine carcinoma (6), Pleiomorphic carcinoma (1), Not specified (4)	Median follow-up: 19.2 months
Ahn SY, 2015 (26020832)	2006–2011	South Korea	Cohort (Seoul National University College of Medicine)	Retrospective	CT texture features (Homogeneity, Kurtosis, Standard deviation, Entropy, Skweness, Mean attenuation of primary tumors)	Clinical stage: IIIA (45), IIIB (53)	TNM7	Concurrent chemoradiotherapy	Squamous cell carcinoma (40), Adenocarcinoma (28), Not specified (30)	NR
Ryu JS, 2014 (24550423)	2002–2010	South Korea	Cohort (Inha University Hospital)	Retrospective	Pleural effusion	Clinical stage: IIIA (227), IIIB (248)	TNM7	Concurrent/sequential chemoradiotherapy (stage IIIA and IIIB), (neoadjuvant chemotherapy and surgery (stage IIIA), or cytotoxic chemotherapy (stage IIIB)	Not specified for stage III; Overall study population: Squamous cell carcinoma (863), Adenocarcinoma (1004), Not specified (194)	NR
Koo TR, 2014 (25498887)	2001–2009	South Korea	Cohort (Seoul National University College of Medicine)	Retrospective	Gross tumor volume, T-stage	NR Stage: IIIA (49), IIIB (108)	TNM7	Concurrent chemoradiotherapy	Squamous cell carcinoma (86), Adenocarcinoma (52), Large cell carcinoma (2), Not specified (17)	Median follow-up: 24.4 months
Gensheimer MF, 2017 (28830717)	2006–2015	USA	Cohort (Stanford University School of Medicine)	Retrospective	Gross tumor volume	NR Stage: IIIA (36), IIIB (41)	NR (TNM6 or TNM7)	Concurrent chemoradiotherapy, or radiotherapy	NR	Median follow-up: 14 months

Fried DV, 2016 (26176655)	2008–2013	USA	Cohort (University of Texas MD Anderson Cancer Center)	Retrospective	T-stage, Gross tumor volume	NR Stage: IIIA (107), IIIB (88)	TNM7	Radiotherapy (and chemotherapy)	Squamous cell carcinoma (89), Not specified (106)	Median follow-up (surviving patients): 37 months
Fried DV, 2014 (25220716)	2004–2012	USA	Cohort (University of Texas MD Anderson Cancer Center)	Retrospective	Gross tumor volume, CT texture features (Average, Kurtosis, Busyness, Infomc1, Standard Deviation, Uniformity, Sosvariance on CE, AVG or T50 CT)	NR Stage: IIIA (45), IIIB (46)	TNM7	Concurrent chemoradiotherapy (and adjuvant chemotherapy)	Squamous cell carcinoma (46), Not specified (45)	Median follow-up (surviving patients): 59 months
Alexander BM, 2011 (20605346)	2000–2006	USA	Cohort (Brigham and Women's Hospital/ Dana- Farber Cancer Institute)	Retrospective	Tumor volume, Nodal volume	NR Stage: IIIA (46), IIIB (61)	TNM6	Concurrent chemoradiotherapy (and surgery)	Squamous cell carcinoma (27), Adenosquamous carcinoma (38), Large cell carcinoma (4), Not specified (38)	Median follow-up: 15 months
Sibley GS, 1995 (7493826)	1987–1992	USA	Cohort (Micheal Reese Hospital)	Retrospective	Tumor volume, Atelectasis, T-stage	NR Stage: IIIA (18), IIIB (19)	TNM4	Radiotherapy (and chemotherapy)	Squamous cell carcinoma (23), Adenocarcinoma (6). Large cell carcinoma (2), Not specified (6)	Median follow-up: 18.9 months
Soussan M, 2013 (23306807)	2009–2011	France	Cohort (Avicenne University Hospital)	Retrospective	Tumor diameter, Tumor volume	Clinical stage: IIIA (23), IIIB (9)	TNM7	Induction chemotherapy and radiotherapy or surgery	Squamous cell carcinoma (16), Adenocarcinoma (12), Large cell carcinoma (4)	Median follow-up: 19 months
Watanabe Y, 2016 (27663793)	1998–2007	Japan	Cohort (National Cancer Center Hospital Tokyo)	Retrospective	Cavitary wall thickness	NR Stage: III (28)	NR (TNM5 or TNM6)	Surgery	Adenocarcinoma (28)	NR
Basaki K, 2006 (16226400)	1997–2003	Japan	Cohort (Hirosaki University Hospital)	Retrospective	T-stage, Gross tumor volume, Tumor volume, Nodal volume, Location	NR Stage: IIIA (30), IIIB (41)	NR (TNM5 or TNM6)	Radiotherapy (and chemotherapy)	Squamous cell carcinoma (56), Adenocarcinoma (12), Large cell carcinoma (2), Adenosquamous carcinoma (1)	Median follow-up: 34 months
Hyun SH, 2014 (23948859)	2003–2007	South Korea	Cohort (Samsung Medical Center)	Retrospective	T-stage	Pathological & clinical stage: IIIA (194)	TNM7	Surgery (and adjuvant chemotherapy and/or radiotherapy)	Squamous cell carcinoma (74), Adenocarcinoma (100), Large cell carcinoma (6), Not specified (14)	Median follow-up: 54 months
William WN, 2009 (19318668)	1998–2003	USA	Cohort SEER database	Retrospective	Tumor diameter	NR Stage: IIIB (22091)	TNM6	Surgery and/or radiotherapy	Squamous cell carcinoma (5725), adenocarcinoma (6841), Large cell carcinoma (1410), Bronchialveolar carcinoma (482), Not specified (6169)	NR
Morgensztern D, 2012 (22982648)	1998–2003	USA	Cohort SEER database	Retrospective	Tumor diameter	NR Stage: IIIA (6327), IIIB (5988)	NR (TNM7)	NR	Squamous cell carcinoma (3920), Adenocarcinoma (3500), Large cell carcinoma (768), Not specified (4127)	Median follow-up: 10 months
Hyun SH 2015 (26295651)	2008–2013	South Korea	Cohort (Samsung Medical Center)	Retrospective	T-stage	Clinical stage: IIIA (161)	NR (TNM6 or TNM7)	Surgery (and adjuvant chemotherapy and/or radiotherapy)	Squamous cell carcinoma (56), Adenocarcinoma (92), Not specified (13)	Median follow-up (surviving patients): 20 months

Table 1. Continued.

7 Study citation (PMID)	Study duration (start – end date)	Country	Source of data	Study design	CT-related prognostic factor	Number of stage III participants	TNM staging system used (estimation)	Treatment of the stage III NSCLC participants	NSCLC histological subtypes present	Outcome measure (follow-up)
Bulbul Y, 2010 (20636252)	2006–2008	Turkey	Cohort (Farabi Hospital)	Prospective	Atelectasis/ Obstructive pneumonitis	NR stage: IIIA (8), IIIB (32)	NR (TNM6 or TNM7)	Sequential chemoradiotherapy	NR	NR
Wald P, 2017 (28843360)	2012–2016	USA	Cohort (The Ohio State University Wexner Medical Center)	Retrospective	T-stage, Gross tumor volume	Clinical stage: IIIA (39), IIIB (13)	NR (TNM7)	Concurrent chemoradiotherapy (and induction/consolidative chemotherapy)	Squamous cell carcinoma (30), Adenocarcinoma (19), Not specified (3)	Median follow-up: 19.3 months
Xiang ZL, 2012 (22929048)	2005 – NR	USA	Cohort (University of Texas MD Anderson Cancer Center)	Retrospective	Gross tumor volume	Clinical stage: III (84)	NR (TNM6 or TNM7)	Concurrent chemoradiotherapy	Squamous cell carcinoma (38), Adenocarcinoma (34), Not specified (12)	Median follow-up: 19.2 months
Wu J, 2016 (27212196)	2005–2009	USA	Cohort (Stanford University School of Medicine)	Retrospective	Gross tumor volume	NR Stage: IIIA (12), IIIB (20)	NR (TNM6 or TNM7)	Radiotherapy (and chemotherapy)	NR	Not specified for stage III, Entire cohort: Median follow-up: 20.2 months
Elsayad K, 2018 (29623466)	2013–2017	Germany	Cohort (University Hospital Münster)	Retrospective	Gross tumor volume	NR stage: IIIA (26), IIIB (13), IIIC (11)	TNM8	Radiotherapy (and chemotherapy)	Squamous cell carcinoma (22), Adenocarcinoma (25), Other (3)	Median follow-up: 10 months
Jie Y, 2017 (NA)	2009–2012	China	Cohort (Shangdong Cancer Hospital)	Retrospective	T-stage, Tumor volume, Location	NR Stage: IIIA (35), IIIB (43)	NR (TNM7)	Concurrent chemoradiotherapy	Squamous cell carcinoma (33), Adenocarcinoma (34), Not specified (11)	Median follow-up: 24.5 months
Shien K, 2015 (NA)	1999–2011	Japan	Cohort (Okayama University Hospital)	Retrospective	Location	Clinical stage: IIIA (44), IIIB (32)	TNM7	Surgery and induction chemoradiotherapy	Squamous cell carcinoma (30), Adenocarcinoma (43), Adenosquamous cell carcinoma (1), Large cell carcinoma (2)	Median follow-up: 64 months
Crvenkova S, 2015 (NA)	2005–2008	Macedonia	Cohort (University Clinic of Radiotherapy and Oncology Skopje)	Prospective	Tumor diameter	NR stage: IIIB (85)	NR (TNM6 or TNM7)	Concurrent/sequential chemoradiotherapy	Squamous cell carcinoma (56), Adenocarcinoma (16), Large cell carcinoma (5), Not specified (8)	Median follow-up: 36 months
Saga T, 2015 (NA)	2010–2014	Japan	Cohort (Cancer Institute Hospital)	Prospective	T-stage	Clinical stage: IIIA (12), IIIB (11)	NR (TNM7)	Concurrent/sequential chemoradiotherapy	Not specified for stage III, Overall study population: Squamous cell carcinoma (11), Adenocarcinoma (19), Large cell carcinoma (8)	NR
Li J, 2009 (NA)	1998–2004	China	Cohort (Hospital of Jiangsu University)	Prospective	T-stage	Clinical stage: IIIA (91)	TNM5	Surgery and neo adjuvant chemotherapy and/or radiotherapy	Squamous cell carcinoma (40), Adenocarcinoma (44), Large cell carcinoma (4), Undifferentiated NSCLC (3)	Median follow-up: 43 months

Dong X, 2016 (NA)	2007–2010	China	Cohort (Shangdong Cancer Hospital)	Retrospective	T-stage, Location	NR stage IIIA (24), IIIB (34)	TNM6	Concurrent chemoradiotherapy	Squamous cell carcinoma (30), Adenocarcinoma (25), Not specified (3)	Median follow-up: 60 months
Agrawal V, 2017 (28426673)	2003–2013	USA	Cohort (Brigham and Women's Hospital/ Dana- Farber Cancer Institute)	Retrospective	T-stage, Gross tumor volume, Tumor volume	NR Stage: IIIA (61), IIIB (12)	TNM7	Concurrent chemoradiotherapy and surgery	Squamous cell carcinoma (16), Adenocarcinoma (48), Not specified (9)	Median follow-up: 36 months
Phernambucq ECJ, 2012 (22659960)	2003–2010	The Netherlands	Cohort (VU University Medical Center)	Retrospective	Tumor cavitation	Clinical stage: IIIA (36), IIIB (51)	NR (TNM6 or TNM7)	Concurrent chemoradiotherapy (and surgery)	Squamous cell carcinoma (37), Adenocarcinoma (28), Large cell carcinoma	NR
Chaft JE, 2013 (23857398)	2005–2011	USA	Cohort (Memorial Sloan- Kettering Cancer Center)	Prospective	Tumor cavitation	Clinical stage: IIIA (34)	TNM6	Neoadjuvant chemoradiotherapy, adjuvant immunotherapy and surgery	Not specified for stage III; Overall study population: Adenocarcinoma (45), Large cell carcinoma (4), Adenosquamous carcinoma (1)	Median follow-up: 29 months
Chang JY, 2017 (28727865)	2009–2011	USA	Cohort (University of Texas MD Anderson Cancer Center)	Prospective	Tumor diameter, Location	NR stage: IIIA (30), IIIB (34)	NR (TNM7)	Concurrent chemoradiotherapy	Squamous cell carcinoma (28), Adenocarcinoma (25), Not specified (11)	Median follow-up: 27.3 months
Naito Y, 2008 (18520801)	2000–2004	Japan	Cohort (National Cancer Center Hospital)	Retrospective	Tumor diameter	Clinical stage: IIIA (26), IIIB (47)	NR (TNM5 or TNM6)	Concurrent chemoradiotherapy	Squamous cell carcinoma (28), Adenocarcinoma (29), Not specified (16)	Median follow-up: 35 months
Shumway, 2011 (21676484)	1999–2010	USA	Cohort (The University of Chicago)	Retrospective	T-stage	Clinical stage: IIIA (44), IIIB (9)	TNM6	Concurrent chemoradiotherapy and surgery	Squamous cell carcinoma (19), Adenocarcinoma (22), Large cell carcinoma (2), Not specified (10)	Median follow-up: 19 months
Nguyen QN, 2015 (26028228)	2006–2010	USA	Cohort (University of Texas MD Anderson Cancer Center)	Prospective	Gross tumor volume	NR stage: IIIA (70), IIIB (43)	TNM6	Concurrent chemoradiotherapy	Not specified for stage III; Overall study population: Squamous cell carcinoma (59), Not specified (75)	Median follow-up: 56.4 months
Etiz D, 2002 (12095548)	1991–1998	USA	Cohort (Duke University Medical Center)	Retrospective	Gross tumor volume	NR stage: IIIA (47), IIIB (64)	TNM5	Concurrent chemoradiotherapy (and induction/adjuvant chemotherapy)	Not specified for stage III; Overall study population: Squamous cell carcinoma (66), Adenocarcinoma (33), Large cell carcinoma (20), Not specified (31)	Median follow-up: 13.2 months
Akcam TI, 2015 (NA)	2005–2011	Turkey	Cohort (Dr Suat Seren Chesr Disease and Surgery Training and Research Hospital)	Retrospective	T-stage, Location	NR stage: IIIA (74), IIIB (37)	TNM7	Surgery and adjuvant chemotherapy	Squamous cell carcinoma (59), Adenocarcinoma (50), Large cell carcinoma (2)	Mean followup: 31.8 months
Zhou R, 2018 (NA)	2005–2013	USA	Cohort (University of Texas MD Anderson Cancer Center)	Retrospective	T-stage, Gross tumor volume	NR stage: IIIA (234), IIIB (257)	TNM6	Concurrent chemoradiotherapy (and induction/adjuvant chemotherapy)	Squamous cell carcinoma (182), Not specified (309)	NR
Park YJ, 2015 (NA)	2009–2011	South Korea	Cohort (Ansan Hospital)	Retrospective	Gross tumor volume	Clinical stage: IIIA (8), IIIB (23)	TNM7	Concurrent chemoradiotherapy	Squamous cell carcinoma (20), Adenocarcinoma (9), Not specified (2)	NR
Oberije C, 2015 (25936599)	2002–2011	The Netherlands	Cohort (MAASTRO clinic)	Prospective	T-stage, Nodal volume	Clinical stage: IIIA (199), IIIB (349)	TNM6	Concurrent/sequential chemoradiotherapy	Squamous cell carcinoma (164), Adenocarcinoma (81), Large cell carcinoma (190), Not specified (113)	Median follow-up: 66 months

Table 1. Continued.

7 Study citation (PMID)	Study duration (start – end date)	Country	Source of data	Study design	CT-related prognostic factor	Number of stage III participants	TNM staging system used (estimation)	Treatment of the stage III NSCLC participants	NSCLC histological subtypes present	Outcome measure (follow-up)
Warner A, 2016 (26867890)	1995–2010	Europe, USA, Asia	Cohort (13 institutions)	Retrospective	Gross tumor volume	NR stage: IIIA (366), IIIB (650), IIINR (143)	NR (TNM5 or TNM6)	Concurrent chemoradiotherapy	Squamous cell carcinoma (338), Adenocarcinoma (289), Large cell carcinoma (145), Not specified (473)	Median follow-up: 43.5 months
Hayakawa K, 1996 (8765179)	1976–1989	Japan	Cohort (Gunma University Hospital)	Retrospective	T-stage, Tumor diameter Location	Clinical stage: IIIA (81), IIIB (60)	TNM4	Radiotherapy (and chemotherapy)	Squamous cell carcinoma (104), Adenocarcinoma (24), Large cell carcinoma (13)	NR
Mao Q, 2018 (29554790)	2004–2009	USA	Cohort SEER database	Retrospective	Tumor diameter, Location	NR stage: IIIA (1809)	TNM7	Surgery and/or radiotherapy	Squamous cell carcinoma (444), Adenocarcinoma (1294), Large cell carcinoma (71)	Median follow-up: 39 months
Pang Z, 2017 (29268415)	2004–2011	USA	Cohort SEER database	Retrospective	Tumor diameter, Location	Clinical stage: IIIA (98700)	TNM7	Surgery and/or radiotherapy	Squamous cell carcinoma (21748), Adenocarcinoma (32175), Not specified (37251)	NR
Broderick SR, 2016 (26410162)	1998–2010	USA	Cohort NCDB	Retrospective	T-stage, Tumor diameter	Clinical stage: IIIA (542)	NR (TNM7)	Surgery and neoadjuvant/adjuvant chemoradiotherapy	NR	NR
Hwang IG, 2008 (18623378)q	1997–2003	South Korea	Cohort (Samsung Medical Center)	Retrospective	T-stage	NR stage: IIIA (68)	TNM5	Surgery and neoadjuvant concurrent chemoradiotherapy	Adenocarcinoma (41), Not specified (27)	Median follow-up: 61.8 months
Topkan E, 2018 (29887509)	2007–2013	Turkey	Cohort (Baskent University Medical Faculty)	Retrospective	Tumor diameter, Tumor cavitation	Clinical stage: IIIA (154), IIIB (635)	TNM7	Concurrent chemoradiotherapy	Squamous cell carcinoma (789)	Median follow-up: 22.9 months
Hishida T, 2014 (24203815)	1993–2008	Japan	Cohort (National Cancer Center Hospital)	Retrospective	Tumor diameter, Location	Clinical stage: IIIA (97)	TNM6	Surgery (and adjuvant chemotherapy and/or radiotherapy)	Squamous cell carcinoma (25), Adenocarcinoma (52), Large cell carcinoma (6), Adenosquamous carcinoma (7), Not specified (7)	Median follow-up: 70.8 months
Horinouchi H, 2012 (23004347)	1999–2003	Japan	Cohort (National Cancer Center Hospital)	Retrospective	T-stage	Clinical stage: IIIA (50), IIIB (61)	NR (TNM5 or TNM6)	Concurrent chemoradiotherapy	Squamous cell carcinoma (26), Adenocarcinoma (71), Large cell carcinoma (6), Adenosquamous carcinoma (1), Not specified (7)	NR
Betticher DC, 2006 (16622435)	1997–2000	Switzerland	Multicenter	Prospective	T-stage, Nodal enlargment	NR stage: IIIA (75)	NR (TNM5)	Surgery and neo adjuvant chemotherapy and/or radiotherapy	Squamous cell carcinoma (32), Adenocarcinoma (23), Large-cell carcinoma (9), Not specified (11)	Median follow-up: 60 months
Kanzaki H, 2016 (27125214)	2006–2012	Japan	Cohort (Shikoku Cancer Center Hospital)	Retrospective	T-stage, Gross tumor volume	Clinical stage: III (111)	TNM7	Concurrent/sequential chemoradiotherapy or radiotherapy	Squamous cell carcinoma (45), Adenocarcinoma (48), Large cell carcinoma (5), Not specified (13)	Median follow-up: 52.2 months
Lee VHF, 2016 (24710123)	2006–2012	Hong Kong	Cohort (Li Ka Shing Faculty of Medicine)	Retrospective	Gross tumor volume, Tumor volume, Nodal volume	NR stage: IIIA (18), IIIB (25)	TNM7	Concurrent chemoradiotherapy (and induction/adjuvant chemotherapy)	Squamous cell carcinoma (9), Adenocarcinoma (25), Not specified (9)	Median follow-up: 41.5 months

Casiraghi M, 2019 (30446406)	1998–2015	Italy	Cohort (European Institute of Oncology, Milan)	Retrospective	Location	Pathological & clinical stage: IIIA (233)	TNM7	Surgery and neo adjuvant chemotherapy and/or radiotherapy	Squamous cell carcinoma (89), Adenocarcinoma (117), Large cell carcinoma (3), Adenosquamous carcinoma (8), Pleomorphic carcinoma (9), Carcinosarcoma (1), Not specified (6)	Median follow-up: 24 months
Higo H, 2019 (30793176)	2012–2015	Japan	Cohort (Okayama University Hospital)	Retrospective	Interstitial lung abnormalities	Clinical stage: III (71)	TNM7	Chemoradiotherapy	Squamous cell carcinoma (25), Adenocarcinoma (40), Not specified (6)	NR
Kim E, 2019 (30266585)	2006–2013	South Korea	Cohort (SMGSNU Boramae Medical Center)	Retrospective	T-stage	Clinical stage: IIIA (72), IIIB (58)	TNM7	Concurrent chemoradiotherapy	Squamous cell carcinoma (64), Adenocarcinoma (44), Not specified (22)	Mean followup: 51.3 months
Dieleman EMT, 2018 (30055239)	2005–2015	The Netherlands	Cohort (AMC)	Retrospective	Tumor volume, Nodal volume	NR stage: IIIA (116), IIIB (38)	TNM7	Concurrent chemoradiotherapy	Squamous cell carcinoma/ Large cell carcinoma (118), Adenocarcinoma (36)	Median follow-up: 22 months
Yoo GS, 2019 (30544255)	1996–2015	South Korea	Cohort (Samsung Medical Center)	Retrospective	Great vessel invasion	NR Stage: IIIA (13), IIIB (24)	NR (TNM7)	Concurrent chemoradiotherapy	Squamous cell carcinoma (21), Adenocarcinoma (11), Not specified (5)	Median follow- up: 17 months
Pusceddu C, 2019 (31289539)	2010–2013	Italy	Cohort (Oncological Hospital A. Businco)	Retrospective	Tumor diameter	NR stage: IIIB/C (53)	NR (TNM8)	Microwave ablation	Squamous cell carcinoma (13), Adenocarcinoma (51), Large cell carcinoma (1)	Median follow-up: 21.5 months
Konert T, 2019, (31367906)	2010–2014		Multicenter	Retrospective	T-stage	NR stage: IIIA (145), IIIB (53), IIIC (32)	NR (TNM8)	Concurrent/sequential chemoradiotherapy or radiotherapy	Squamous cell carcinoma (90), Adenocarcinoma (97), Large cell carcinoma (15), Not specified (28)	Median follow-up: 15 months
Maniwa T, 2018 (30746228)	2006–2013	Japan	Cohort (12 thoracic surgery departments belonging to the Thoracic Surgery Study Group of Osaka University)	Retrospective	T-stage, Tumor diameter	Clinical stage: IIIA (92), IIIB (2)	TNM7	Surgery and adjuvant chemotherapy	Adenocarcinoma (65), Not specified (29)	Median follow-up: 56.5 months
Tao X, 2019 (31179087)	2007–2016	China	Cohort (Fudan University Shanghai Cancer Center)	Retrospective	T-stage, Location	Pathological & clinical stage: IIIA (603)	TNM8	Surgery and neoadjuvant/adjuvant chemoradiotherapy	Squamous cell carcinoma (135), Adenocarcinoma (425), Adenosquamous carcinoma (26), Not specified (17)	Median follow-up: 31.98 months
Kim DY, 2019 (31591865)	2004–2016	South Korea	Cohort (Seoul National University Bundang Hospital)	Retrospective	T-stage	NR stage IIIA (56), IIIB (26)	TNM7	Chemoradiotherapy	Squamous cell carcinoma (52), Adenocarcinoma (16), Not specified (14)	Median follow-up (surviving patients): 20.1 months

Table 2. Summary of findings radiomics and other CT-related prognostic factors.

Radiomic feature related prognostic factor	**Study citation (First author, year (PMID))**	**Description texture measurement**	**Description prognostic factor groups**	**Number of patients**	**OS**	**Univariate analysis (estimate (95% CI))**	**Multivariable analysis (estimate (95% CI)**	**Factors corrected for in multivariable analysis**
Homogeneity	Ahn SY, 2015 (26020832)	CE-CT	>0.03 vs ≤0.03	54/44	Mean (months): 24.8/ Median (months): 23.0	$p = 0.483$		
Kurtosis	Ahn SY, 2015 (26020832)	CE-CT	>9.932 vs ≤9.932	49/49	Mean (months): 25.3/ Median (months): 21.0	$p = 0.488$		
	Fried DV, 2014 (25220716)	CE-CT; IHIST T50-CT; GRAD	Continuous Continuous	91 91			HR: 0.978 Not included in model	
Standard deviation	Ahn SY, 2015 (26020832)	CE-CT	>36.411 vs ≤36.411	43/55	Mean (months): 26.0/ Median (months): 21.0	$p = 0.295$		
	Fried DV, 2014 (25220716)	AVG-CT; LoG	Continuous	91			HR: 1.024	
Entropy	Ahn SY, 2015 (26020832)	CE-CT	≤4.445 vs >4.445	23/75	Mean (months): 29.8/ Median (months): 20.0	$p = 0.030$	HR: 2.31 (1.031–5.226) $p = 0.040$	Skewness, Mean HU
Skewness	Ahn SY, 2015 (26020832)	CE-CT	≤—2.374 vs >—2.374	38/60	Mean (months): 28.1/ Median (months): 19.0	$p = 0.021$	HR: 1.92 (1.013–3.642) $p = 0.046$	Entropy, Mean HU
Mean HU	Ahn SY, 2015 (26020832)	CE-CT	≤43.448 vs >43.448	49/49	Mean (months):26.8/ Median (months): 17.0	$p = 0.030$	HR: 1.93 (1.074–3.454) $p = 0.028$	Entropy, Skewness
Largest axial slice average	Fried DV, 2014 (25220716)	CE-CT; LoG; Sigma = 1 T50-CT; LoG; sigma = 1.5	Continuous Continuous	91 91			HR: 1.15 HR: 0.923	
Average	Fried DV, 2014 (25220716)	CE-CT; LoG; Sigma = 1	Continuous	91			Not included in model	
Largest axial slice uniformity	Fried DV, 2014 (25220716)	AVG-CT; LoG; Sigma = 1 AVG-CT; LoG; sigma = 2.5 T50-CT; LoG; Sigma = 1.5	Continuous Continuous Continuous	91 91 91			HR: 1.54 HR: 1.73 Not included in model	
Busyness	Fried DV, 2014 (25220716)	CE-CT; NGTDM	Continuous	91			Not included in model	
Infomc1	Fried DV, 2014 (25220716)	CE-CT, COM	Continuous	91			HR: 12.2	
Sosvariance	Fried DV, 2014 (25220716)	T50-CT; COM	Continuous	91			HR: 1.0011	
Other CT-related prognostic factor	Study citation (First author, year (PMID))	Description prognostic factor groups	Number of patients	OS	Univariate analysis (estimate (95% CI))	Multivariable analysis (estimate (95% CI)	Factors corrected for in multivariable analysis	
Atelectasis/ Obstructive pneumonitis	Sibley GS, 1995 (7493826)	None vs <50% vs >50%	21/5/11.	Median (months): 18.3/19.5/19.8 1y survival rate (%): 61/100/91 2y survival rate (%): 45/0/28	$p = 0.98$			
	Bulbul Y, 2010 (20636252)	Negative vs Positive (Stage IIIA & IIIB patients)		Median (months): 14.5/9.8 1y survival rate (%): 67.2/40.0	$p = 0.032$			
		Negative vs Positive (Stage IIIB patients)		Median (months): 13.9/9.3 1y survival rate (%): 72.2/35.8	$p = 0.044$			
Interstitial lung abnormalities	Higo H, 2019 (30793176)	Positive vs Negative			$p = 0.49$			

Location	Dong X, 2016 (27322376)	Right vs Left	40/18		HR: 1.756 (0.718–1.958) p = 0.637		
	Basaki K, 2006 (16226400)	Right vs Left	41/30	Median (months): 14/12 2y survival rate (%): 18/29	Not significant		
		Hilar vs Upper vs Middlelower	33/29/9	Median (months): 15/12/9 2y survival rate (%): 29/20/0	Not significant		
	Pang Z, 2017 (29268415)	Left vs Right	35946/58435		HR: 1.006 (0.992–1.020) p = 0.406		
	Casiraghi M, 2019 (30446406)	Right vs Left	130/103		HR: 0.98 (0.72– 1.33) p = 0.89		
	Jie Y, 2017 (NA)	Central vs Peripheral	37/41		HR: 1.464 (0.871–2.463) p = 0.151		
	Tao X, 2019 (31179087)	Central vs Peripheral	128/475		HR: 1.08 (0.74–1.57) p = 0.6843		
	Shien K, 2015 (NA)	Non-lower lobe vs Lower lobe	58/18	5y survival rate (%): 77.0/37.9	p = 0.022		
	Chang JY, 2017 (28727865)	Left lung or right lower lobe vs Right middle or right upper lobe				HR: 1.90 (1.03–3.50) p = 0.04	KPS, Overall stage, Tumor size
	Akcam TI, 2015 (NA)	Upper lobe vs Middle lobe vs Lower lobe	64/3/44			HR: 1.538 (0.968–2.445) p = 0.069	Age, Histology, T-stage, Multi-single station
	Hayakawa K, 1996 (8765179)	Upper lobe vs Superior segment of the lower lobe vs Lower lobe vs Main or intermediate bronchus	83/19/28/11	Median (months): 13.5/16/12/9.5 2y survival rate (%): 25/42/12/0 5y survival rate (%): 16/5/4/0	p = 0.032		
		Upper lobe + Superior segment of the lower lobe vs Lower lobe	102/28			HR: 1.51 (1.12–2.04) p = 0.0085	Age, Gender, PS, Histology, Tumor size, T-stage, Nstage, Total dose, Field size
		Upper lobe + Superior segment of the lower lobe vs Main or intermediate bronchus	102/11			HR: 2.28 (1.24–4.16) p = 0.0085	Age, Gender, PS, Histology, Tumor size, T-stage, Nstage, Total dose, Field size
	Mao Q, 2018 (29554790)	Main bronchus vs Upper lobe	22/1043	Median (months): 36.0/40.0	HR: 0.856 (0.427–1.714) p = 0.660		
		Main bronchus vs Middle lobe	22/84	Median (months): 36.0/42.0	HR: 0.697 (0.418–1.162) p = 0.167		
		Main bronchus vs Lower lobe	22/602	Median (months): 36.0/34.0	HR: 0.665 (0.374–1.181) p = 0.164		
		Main bronchus vs Overlap lobe	22/39	Median (months): 36.0/28.0	HR: 0.790 (0.472–1.321) p = 0.368		
	Hishida T, 2014 (24203815)	Upper lobe vs Middle or lower lobe	29/16	5y survival rate (%): 29.2/12.5	p = 0.208		
Pit-fall sign	Li M, 2004 (15541820)	Negative vs Positive (Stage III patients)	10/6.	5y survival rate (%): 25.0/50.0	p = 0.470		
		Negative vs Positive (Stage IIIA patients)	7/3.	5y survival rate (%): 14.3/33.3	p = 0.579		
		Negative vs Positive (Stage IIIA patients)	3/3.	5y survival rate (%): 66.7/66.7	p = 0.886		

Table 2. Continued.

Radiomic feature related prognostic factor	Study citation (First author, year (PMID))	Description texture measurement	Description prognostic factor groups	Number of patients	OS	Univariate analysis (estimate (95% CI))	Multivariable analysis (estimate (95% CI)	Factors corrected for in multivariable analysis
Pleural effusion	Ryu JS, 2014 (24550423)	No pleural effusion vs Minimal pleural effusion (Stage IIIA patients)	197/30	Median (months): 17.7/10.6	HR: 2.12 (1.39–3.23) $p = 0.0003$	HR: 1.62 (0.95–2.94)	Gender, Age, Smoking habit, CCI score, ECOG, Weight loss, Hemoglobin, Albumin, Alkaline phosphatase, Calcium, Histology, EGFR mutation, Tumor size, N stage, Number of organs effected by metastasis, PET, Treatment	
		No pleural effusion vs Minimal pleural effusion (Stage IIIB patients)	189/59	Median (months): 14.5/7.8	HR: 1.65 (1.22–2.21) $p < 0.0001$	HR: 1.57 (1.08–2.28)	Gender, Age, Smoking habit, CCI score, ECOG, Weight loss, Hemoglobin, Albumin, Alkaline phosphatase, Calcium, Histology, EGFR mutation, Tumor size, N stage, Number of organs effected by metastasis, PET, Treatment	
Cavitary wall thickness	Watanabe Y, 2016 (27663793)	≤4.5 mm (‘thin’) vs >4.5 mm (’thick’)	7/21.		$p = 0.96$			
Cavitation	Phernambucq ECJ, 2012 (22659960)	Positive vs negative	16/71.	Median (months): 9.9/16.3	$p = 0.09$			
	Chaft JE, 2013 (23857398)	Positive vs negative		3y survival rate (%): 57/44	$p = 0.48$			
	Topkan E, 2018 (29887509)	Positive vs negative	694/95.	Median (months): 24.1/15.7	$p < 0.001$	HR: 1.54 (1.37–1.71) $p < 0.001$	Overall stage, Weight loss status, Anemia	
Great vessel invasion	Yoo GS, 2019 (30544255)	Aortic arch	4	2y survival rate (%): 75.0	$p = 0.065$	HR: 0.058 (0.002–2.25) $p = 0.127$	Age, Gender, PS, Histology, N-stage	
		Descending aorta	3	2y survival rate (%): 33.3	$p = 0.189$	HR: 3.60 (0.30–43.02) $p = 0.312$	Age, Gender, PS, Histology, N-stage	
		Pulmonary artery	13	2y survival rate (%): 51.9	$p = 0.883$	HR: 0.53 (0.074–3.73) $p = 0.520$	Age, Gender, PS, Histology, N-stage	
		Superior vena cava	10	2y survival rate (%): 62.5	$p = 0.579$	HR: 0.16 (0.008–3.31) $p = 0.235$	Age, Gender, PS, Histology, N-stage	
		Heart	11	2y survival rate (%): 24.5	$p = 0.218$	HR: 1.94 (0.24–15.75) $p = 0.537$	Age, Gender, PS, Histology, N-stage	

Outcomes concerning radiomic features and other CT-related prognostic factors in univariate and multivariable analysis of the included studies. Estimates are reported with 95% confidence interval and p-value when available. Used statistical models are: Cox proportional hazard model, log-rank test (LR). Abbreviations: AVG: Average intensity projection image, CCI: Charlson comorbidity index, CE-CT: Contrast enhanced computed tomography, CI: Confidence interval, COM: Co-occurrence matrix, ECOG: Eastern cooperative oncology group, EGFR: Epidermal growth factor receptor, GRAD: Absolute gradient, HR: Hazard ratio, HU: Hounsfield unit, IHIST: Histogram, LoG: Laplacian of Gaussian filter, NGTDM: Nearest gray tone difference matrix, OS: Overall survival, PET: Positron emission tomography, PMID: PubMed identification number, T50: Expiratory image.

Table 3. Summary of findings size-related prognostic factors.

Size-related prognostic factors	Study citation (First author, year (PMID))	Description prognostic factor groups	Number of patients	OS	Univariate analysis (estimate (95% CI))	Multivariable analysis (estimate (95% CI)	Factors corrected for in multivariable analysis
Gross tumor volume (GTV)	Koo TR, 2014 (25498887)	≤50 cm3 vs >50 cm^3	33/124	3y survival rate (%): 65.7/28.4	$p < 0.001$		
		Continuous		Median (months): 25.5; 3y survival rate (%): 36.4	HR: 1.001 (1.000–1.002) $p = 0.019$		
	Basaki K, 2006 (16226400)	<85 mL vs >85 mL	36/35	Median (months): 18/11; 2y survival rate (%): 34/10	$p = 0.0003$	HR: 1.05 (1.02–1.09) $p < 0.01$	PS, 2 Gy equivalent dose, Age, Chemotherapy, Histology, Tstage, N-stage
	Xiang ZL, 2012 (22929048)	<96.6 cm^3 vs ≥96.6 cm^3	42/42		HR: 1.764 (0.866–3.592) $p = 0.118$		
	Etiz D, 2002 (12095548)	<97 cm^3 vs ≥97 cm^3				$p = 0.006$	Age, Gender, KPS, Weight loss, Nstage, Total dose (6 Gy), Fractionation schedule, Chemotherapy
	Park YJ, 2015 (NA)	≥90 cm3 vs <90 cm3		Median (months); 15.8/13.0	$p = 0.670$		
	Warner A, 2016 (26867890)	≥100 cm3 vs <100 cm3			OR: 2.53 (1.53–4.18) $p < 0.001$	OR: 2.61 (1.10–6.20) $p = 0.029$	FEV

		Continuous (50 cm^3)	1245	Median (months): 20.94;
				1y survival rate (%): 70.6%;
				2y survival rate (%): 45.1;
				3y survival rate (%): 31.5;
				4y survival rate (%): 26.8;
				5y survival rate (%): 22.0
					OR: 1.08 (1.00–1.17)
					$p = 0.053$
						OR: 1.04 (0.93–1.17)
						$p = 0.475$
							FEV

	Wu J, 2016 (27212196)	<median vs > median	16/16		HR: 2.75 (1.13–6.72) $p = 0.020$	HR: 1.00 (0.99–1.00) $p = 0.410$	High-risk tumor volume, Overall stage, KPS
	Gensheimer MF, 2017 (28830717)	Continuous	77	Median (months): 23; 2y survival rate: 46%	HR: 1.33 (0.94–1.90) $p = 0.110$		
	Fried DV, 2016 (26176655)	Continuous	195			HR: 1.252, $p = 0.01$	Overall stage, T-stage, Induction chemotherapy, Age, Gender, KPS, Co-occurance matrix energy, Solidity
	Fried DV, 2014 (25220716)	Continuous	91			HR: 1.0024	Age, ECOG, Histology, Gender, Texture features (Average, Kurtosis, Busyness, Infomc1, Standard Deviation, Uniformity, Sosvariance on CE, AVG or T50 CT)
	Wald P, 2017 (28843360)	Continuous	53	2y survival rate (%): 53.9	HR: 1.00 (1.00–1.01) $p = 0.983$		
	Elsayad K, 2018 (29623466)	Continuous	50	Median (months): 20; 2y survival rate (%): 46	HR: 1.002 (1–1.004), $p = 0.06$		

	Agrawal V, 2017 (28426673)	Continuous	73	Median (months): 78;
				1y survival rate (%): 85;
				3y survival rate (%): 68
					HR: 1.00 (0.99–1.00)
					$p = 0.72$		

	Nguyen QN, 2015 (26028228)	Continuous	113	Median (months): 30.4	HR: 1.437 (1.531–1.7918) $p = 0.00124$	HR: 1.474 (1.177–1.845) $p = 0.007$	Age

Table 3. Continued.

Size-related prognostic factors	Study citation (First author, year (PMID))	Description prognostic factor groups	Number of patients	OS	Univariate analysis (estimate (95% CI))	Multivariable analysis (estimate (95% CI)	Factors corrected for in multivariable analysis

	Zhou R, 2018 (NA)	Continuous	491	Median (months): 21;
				1y survival rate (%): 85.5;
				2y survival rate (%): 61.2;
				3y survival rate (%): 44.5;
				4y survival rate (%): 37.0;
				5y survival rate (%): 31.6
					HR: 1.00 (1.000–1.004)
					$p = 0.042$		

	Kanzaki H, 2016 (27125214)	Continuous (10 mL)	111	Median (months): 21.7; 5y survival rate (%): 22.6	HR: 1.02 (1.00–1.04) $p = 0.013$		
	Lee VHF, 2016 (24710123)	Continuous	43	Median (months): 37.8	$p = 0.059$	$p = 0.049$	Stage, SUV, Nodal volume

Tumor volume	Sibley GS, 1995 (7493826)	<100 cm^3 vs 100–200 cm3 vs >200 cm^3	6/7/8	Median (months): 41/11.3/25.5;
				1y survival rate (%): 67/43/75;
				2y overall survival rate (%): 67/29/49
					$p = 0.55$		

	Basaki K, 2006 (16226400)	<52 cm^3 vs >52 cm^3	36/35	Median (months): 18/10; 2y survival rate (%): 34/9	$p = 0.00008$	HR: 1.05 (1.02–1.09) $p < 0.01$	Nodal volume, PS, 2 Gy equivalent dose, Age, Chemotherapy, Histology, Tstage, N-stage
	Jie Y, 2017 (NA)	<52 cm^3 vs ≥52 cm^3	39/39		HR: 0.667 (0.393–1.131) $p = 0.133$	HR: 0.633 (0.357–1.124) $p = 0.118$	T-stage, N-stage AUC CSH, SUVmax, MTV, TLG
	Soussan M, 2013 (23306807)	Continuous	32	Median (months): 18	NS		
	Alexander BM, 2011 (20605346)	Continuous (by 10 cm^3 increase) – All participants	107	Median (months): 23	HR: 1.01 $p = 0.47$		
		Continuous (by 10 cm^3 increase) – Only chemoradiation paricipants	76	Median (months): 15	HR: 1.02 $p = 0.16$	HR: 1.03 (1.01–1.06) $p < 0.01$	Gender, Nodal volume

	Agrawal V, 2017 (28426673)	Continuous	73	Median (months): 78;
				1y survival rate (%): 85;
				3y survival rate (%): 68
					HR: 1.00 (0.99–1.00)
					$p = 0.52$		

	Lee VHF, 2016 (24710123)	Continuous	43	Median (months): 37.8	$p = 0.064$	$p = 0.069$	GTV, Stage, SUV
	Dieleman EMT, 2018 (30055239)	Continuous	154	Median (months): 36.1; 1y survival rate (%): 79, 2y survival rate (%): 61, 3y survival rate (%): 52, 5y survival rate (%): 40	HR: 1.001 (0.999–1.002) $p = 0.27$		
Tumor diameter	Huo X, 2017 (29441096)	<3.0 cm vs 3.0–5.0 cm vs 5.1–7.0 cm	62/37/83	1y survival rate (%): 93.54/83.78/72.93; 3y survival rate (%): 42.64/23.45/11.19; 5y survival rate (%): 17.50/4.47/0	$p < 0.001$		
	Firat S, 2002 (12243808)	<7cm vs ≥7 cm	47/37		$p = 0.16$		
	Lee HY, 2012 (22265854)	≤4.2 cm vs >4.2 cm			HR: 0.95 (0.57–1.59) $p = 0.844$		

Crvenkova S, 2015 (NA)	≤5cm vs >5 cm	32/47	Median (months): 20/13	$p < 0.001$		

William WN, 2009 (19318668)	<4.5 cm vs >4.5 cm (T4 Satelite patients)	1495/544	Median (months): 27/11;
			2y survival rate (%): 52/24;
			5y survival rate (%): 31/14
					HR: 1.52 (1.32–1.75)
					$p < 0.001$
						Age, Gender, Ethicity, Histology, N-stage, Initial treatment modality

	<4.5 cm vs >4.5 cm (T4 Invasive patients)	2256/3758	Median (months): 12/10;
			2y survival rate (%): 28/20;
			5y survival rate (%): 12/9
					HR: 1.24 (1.16–1.32)
					$p < 0.001$
						Age, Gender, Ethicity, Histology, N-stage, Initial treatment modality

	<4.5 cm vs >4.5 cm (T4 Pleural effusion patients)	2651/2454	Median (months): 6/4;
			2y survival rate (%): 15/9;
			5y survival rate (%): 3/2
					HR: 1.29 (1.21–1.38)
					$p < 0.001$
						Age, Gender, Ethicity, Histology, N-stage, Initial treatment modality

Morgensztern D, 2012 (22982648)	0.1–3.0 cm vs 3.1–5 cm	3499/4245		HR: 1.13 (1.08–1.18) $p < 0.001$		
	0.1–3.0 cm vs 5.1–7 cm	4245/2646		HR: 1.27 (1.21–1.34) $p < 0.001$		
	0.1–3.0 cm vs 7.1–20 cm	2646/1926		HR: 1.41 (1.33–1.50) $p < 0.001$		

	0.1–3.0 cm vs 3.1–5 cm (Stage IIIA patients)		Median (months): 13/11;
			1y survival rate (%): 50.3/43.9;
			2y survival rate (%): 27.7/22.4;
			3y survival rate (%): 17.1/13.2;
			5y survival rate (%): 8.6/7.2
					HR: 1.11 (1.04–1.19)
					$p = 0.001$
						Age, Gender, Ethnicity, Histology, Overall stage

	0.1–3.0 cm-5.1–7 cm (Stage IIIA patients)		Median (months): 11/9;
			1y survival rate (%): 43.9/38.9;
			2y survival rate (%): 22.4/16.5;
			3y survival rate (%): 13.2/11.3;
			5y survival rate (%): 7.2/5.6
					HR: 1.15 (1.07–1.24)
					$p = 0.0001$
						Age, Gender, Ethnicity, Histology, Overall stage

	0.1–3.0 cm vs 7.1–20 cm (Stage IIIA patients)		Median (months): 9/8;
			1y survival rate (%): 38.9/35.2;
			2y survival rate (%): 16.5/14.5;
			3y survival rate (%): 11.3/9.1;
			5y survival rate (%): 5.6/3.7
					HR: 1.15 (1.05–1.26)
					$p = 0.002$
						Age, Gender, Ethnicity, Histology, Overall stage

	0.1–3.0 cm vs 3.1–5 cm (Stage IIIB patients)		Median (months): 11/10;
			1y survival rate (%): 43.8/40.1;
			2y survival rate (%): 22.4/18.7;
			3y survival rate (%): 13.3/10.0;
			5y survival rate (%): 8.3/4.9
					HR: 1.09 (1.01–1.19)
					$p = 0.02$
						Age, Gender, Ethnicity, Histology, Overall stage

	0.1–3.0 cm vs 5.1–7 cm (Stage IIIB patients)		Median (months): 10/9;
			1y survival rate (%): 40.1/35.1;
			2y survival rate (%): 18.7/15.4;
			3y survival rate (%): 10.0/9.3;
			5y survival rate (%): 4.9/5.7
					HR: 1.11 (1.04–1.20)
					$p = 0.003$
						Age, Gender, Ethnicity, Histology, Overall stage

	0.1–3.0 cm vs 7.1–20 cm (Stage IIIB patients)		Median (months): 9/8;
			1y survival rate (%): 35.1/31.6;
			2y survival rate (%): 15.4/13.6;
			3y survival rate (%): 9.3/8.7;
			5y survival rate (%): 5.7/5.4
					HR: 1.10 (1.01–1.19)
					$p = 0.02$
						Age, Gender, Ethnicity, Histology, Overall stage

Table 3. Continued.

Size-related prognostic factors	Study citation (First author, year (PMID))	Description prognostic factor groups	Number of patients	OS	Univariate analysis (estimate (95% CI))	Multivariable analysis (estimate (95% CI)	Factors corrected for in multivariable analysis
	Chang JY, 2017 (28727865)	≤7cm vs >7 cm				HR: 2.39 (1.07–5.31) $p = 0.03$	KPS, Overall stage, Tumor location
	Naito Y, 2008 (18520801)	<5cm vs ≥5 cm	33/40			HR: 0.862 (0.473–1.569) $p = 0.626$	Age, Gender, PS, Overall stage, Smoking status, Histology, Body weight loss

	Hayakawa K, 1996 (8765179)	≤5cm vs >5 cm	44/97	Median (months): 18.5/11.5;
				2y survival rate (%): 35/18;
				5y survival rate (%): 19/7
					$p = 0.008$	HR: 1.41 (0.93–2.14) $p = 0.10$	Age, Gender, PS, Histology, Tstage, N-stage, Location, Total dose, Field size

	Pang Z, 2017 (29268415)	≤3 cm vs 3–5 cm			HR: 1.184 (1.161–1.207) $p = 0.009$	HR: 1.115 (1.093–1.136) $p < 0.001$	Age, Gender, Histology, Location, Differentiation, Surgery type, Therapy
		≤3 cm vs 5–7 cm			HR: 1.332 (1.304–1.361) $p < 0.001$	HR: 1.256 (1.228–1.283) $p < 0.001$	Age, Gender, Histology, Location, Differentiation, Surgery type, Therapy
		≤3 cm vs >7 cm			HR: 1.476 (1.680–1.745) $p < 0.001$	HR: 1.361 (1.329–1.394) $p < 0.001$	Age, Gender, Histology, Location, Differentiation, Surgery type, Therapy
	Topkan E, 2018 (29887509)	≤5 cm vs >5 cm	246/543	Median (months): 25.3/22.4	$p = 0.04$		
	Hishida T, 2014 (24203815)	≤3 cm vs >3 cm	7/38.	5y survival rate (%): 68.6/15.0	$p = 0.106$		
	Pusceddu C, 2019 (31289539)	<4 cm vs ≥4 cm	26/39		$p = 0.03$		
	Maniwa T, 2018 (30746228)	≤3 cm vs >3 cm				HR: 1.42 (0.73–2.87) $p = 0.31$	Single/Multiple N2, Histology
	Soussan M, 2013 (23306807)	Continuous	32	Median: 18 months	NS		
	Mao Q, 2018 (29554790)	Continuous (mm)	1809		HR: 1.010 (1.008–1.012) $p < 0.001$	HR: 1.011 (1.008–1.014) $p < 0.001$	Age, Gender, Location, Histology, Grade, Lymph node number, Positive lymph nodes, Visceral pleural invasion, Surgery type, Therapy
	Broderick SR, 2016 (26410162)	Continuous	542			HR: 1.00 (0.997–1.002) $p = 0.836$	Age, Gender, Ethnicity, Income, Facility type, T-stage, Charlson/ Deyo score, Neoadjuvant therapy, Right pneumonectomy

Outcomes concerning size-related prognostic factors in univariate and multivariable analysis of the included studies. Estimates are reported with 95% confidence interval and p-value when available. Used statistical models and tests are: Cox proportional hazard model, logistic regression, log-rank test (LR), binary proportion test. Abbreviations: AUC-CSH: Area under the curve of cumulative SUV histograms, AVG: Average intensity projection images, CE: Contrast enhanced, CI: Confidence interval, ECOG: Eastern cooperative oncology group, GTV: Gross tumor volume, Gy: Gray, HR: Hazard ratio, (K)PS: (Karnofsky) performance status, MTV: Metabolic tumor volume, OR: odds ratio, OS: Overall survival, PMID: PubMed identification number, SD: Standard deviation, SUV: Standardized uptake volume, T50: Expiratory phase images, TLG: Total lesion glycolysis, y: Year.

Tumor diameter (the longest diameter of the primary tumor in the transverse plane) was tested as a prognostic factor for OS in 17 inclusions [16,32,34,38,45,48–52,57,58,60,68,69,78,83]. These studies had diverse characteristics. Of the 5 studies that consisted solely of stage IIIA patients receiving surgery, 3 did not find significance in univariate and multivariable analysis [38,49,50,52,57,78]. The 2 studies that did find significance were derived from the SEER database with a similar inclusion period, meaning they may contain overlapping data. Taking this into account, the majority of included analyses indicate tumor diameter is not prognostic for this patient subgroup. Significance was reported in 3 of the 6 studies consisting of stage IIIA and IIIB patients receiving chemoradiation [34,45,48,51,58,60]. However 2 of the studies that respectively reported insignificance in univariate and multivariate analysis used an older version of the TNM-staging system (TNM2/3 and TNM4) and were therefore less comparable with the other studies. This implies tumor diameter is a prognostic factor for stage IIIA/B patients treated with chemoradiation. Three studies consisting of stage IIIA and IIIB patients that did not specify treatment [69] or included surgery as treatment modality [16,57], decreasing their comparability to the other 6, respectively reported significance in univariate and multivariable analysis, and insignificance in univariate and multivariate analyses. The final 3 studies found significance in univariate and multivariable analysis, consisting exclusively of stage IIIB/C patients, who received chemoradiation [32], surgery [68], or microwave ablation in their respective studies [83]. While Morgensztern et al. (2012) [69] also did a subgroup analysis for stage IIIB patients, it should be taken into consideration that both William et al. (2009) [68] and Morgensztern et al. (2012) [69] extracted data from the SEER database using the same inclusion period, and are therefore likely to have overlapping data. Therefore, included data indicates tumor diameter is prognostic for stage IIIB NSCLC patients, as all included analyses indicated significance.

Tumor volume was studied in 8 publications [16,18,66,67,71,73,80,81], which were relatively comparable, with exception of 3 studies, including Alexander et al. (2011) [66], which consisted of cohorts where surgery was a treatment option [16,66,73]. In these 3 studies, 2 of which were conducted at the same institution with overlapping inclusion period, tumor volume was insignificant in univariate analysis [16,66,73]. However, Alexander et al. (2011) [66] did a subgroup analysis for patients receiving only chemoradiation, which was comparable in characteristics to the other 5 studies [18,67,71,80,81]. In these studies significance was reported in 1 out of 5 univariate [18], and 2 out of 4 multivariable analyses [18,66]. Nevertheless, it should be taken into consideration that one of the studies which reported insignificance made use of version 4 of the TNM staging system, and was therefore perceived as less relevant in data analysis. Therefore included data is too heterogeneous to make firm conclusions regarding tumor volume as a prognostic factor for stage III NSCLC patients receiving chemoradiation.

The prognostic effect of GTV was studied in 17 inclusions [11,13,15,18,21,35,41,47,54,64,65,70,73,74,76,77,80]. GTV was significant in 8 out of 16 univariate and 7 out of 9 multivariable analyses. It should, however, be taken into consideration that 1 study, which reported insignificance in univariate analysis, had surgery as a treatment option [73], complicating its comparison with other inclusions. Other than this, the cohorts of included studies seemed to correspond concerning treatment and composition of stage. Two publications that reported significance and insignificance in univariate analysis respectively were conducted at the same institution with a similar recruitment period [64,70]. Chance of overlapping data was also present in 5 other studies [11,35,41,65,76], 4 of which reported significance in univariate and multivariable analyses, and 1 insignificance in univariate analysis.

The univariate results of eligible inclusions for GTV were pooled in a meta-analysis (Fig. 2A). Warner et al. (2016) [77] was excluded as it reported an Odds Ratio from a logistic regression, as opposed to a HR. Three studies were excluded from the meta-analysis as the reported point estimate of the HR coincided numerically with either the upper or lower bound on the confidence interval [21,73,76]. Lee et al. (2016) [80] was excluded as it did not report the point estimate or the confidence interval. The five remaining inclusions had no reason to suspect overlapping patient cohorts [13,15,35,54,64]. None of these studies included surgery as a treatment option. Three inclusions did not report the unit of measurement for GTV [13,15,64]. For these studies, the unit of measurement was inferred from the reported median or mean tumor volume. The estimated heterogeneity between these studies was substantial ($I^2 = 50.2\%$, $\tau = 0.12$). The pooled estimate for the HR of GTV measured in units of 100 cm^3 for overall survival is HR = 1.22 (95% CI 1.05–1.42, $p = 0.008$). Considering this evidence, along with the observation that it was significant in majority of multivariable analyses even when comparability of the studies and potential overlapping data was taken into account, it is likely that Gross Tumor Volume is a prognostic factor.

In the 29 publications studying clinical T-stage, T-stage was divided in several different discrete groups (Table 4) [15,18,21,31,36,37,40,43,46,48,50,53,54,56–59,61,62,65,67,71–73,75,76,79,84,85]. Fourteen publications evaluated its prognostic influence dichotomized in a T3- and T4-stage group and a (T0-) T1- and T2-stage group [15,21,31,48,53,54,56,58,59,61,65,71,72,85]. In this way, T-stage was found to be significant in univariate analysis of 2 studies [54,58] and multivariable analysis of 1 study [48], but insignificant in univariate and multivariable analyses of the other 11 studies [15,21,31,53,56,59,61,65,71,72,85]. However, characteristics of 1 study, which did not report significance in univariate analysis, were different considering study population, consisting of clinically and pathologically staged IIIA patients, and treatment, including surgery. Therefore, these studies cannot be directly compared with the other studies, which were comparable regarding study characteristics (Table 1) [61]. It should also be taken into account that the studies used different versions of the TNM staging system, which considering the changes made in T-stage between TNM6/7/8 further complicates the comparison of the studies.

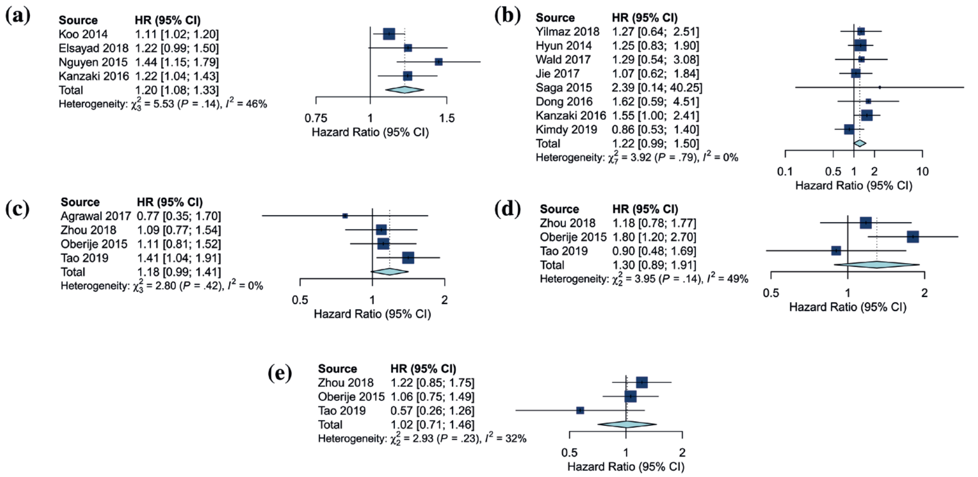


Fig. 2. Forest plot outcome meta-analysis: Forest plots of the outcome of the meta-analysis of: (A) GTV, (B) T1-2 vs T3-4, (C)T1 vs T2, (D) T1 vs T3, and (E) T1 vs T4. In (A) while the results from Gensheimer et al. [64] were included in the meta-analysis, the weight of the study was 0.0% due to the high variance. We excluded these results from the forest plot because they made visual comparison of the other studies impossible.

In order to pool the reported results, the presence of T0 patients was ignored and T0–2 was assumed to be equivalent to T1–2, as the proportion of T0 patients was <2% [54]: Eight inclusions reported a HR and confidence interval for the T-stage 1–2 vs 3–4 comparison with a total of 677 patients (Fig. 2B) [21,31,54,59,61,71,72,85]. There was no indication of heterogeneity between the studies (I^2 = 0.00%, τ = 0.00). The pooled HR was 1.22 (95% CI 0.99–1.50, *p*-value = 0.06). This result is close to the nominal statistical significance level. As a sensitivity analysis we performed a meta-analysis excluding studies that reported surgery as a treatment option (7 studies, pooled HR 1.21, 95% CI 0.95–1.53, *p* = 0.12), and restricting to TNM 7 studies (6 studies, pooled HR 1.20, 95% CI 0.96–1.50, *p* = 0.11), leading to similar results. Taking all this into account, it is unlikely that the univariate clinical T1– 2 vs 3–4 comparison holds prognostic value within a stage III cohort. Also, the majority of comparable multivariable analyses found no significant correlation with OS.

Secondly, 8 studies compared T1-stage with T2-stage, T3-stage, and T4-stage. This comparison did not yield significance in any of the reported univariate and multivariable analyses [18,36,48,57,62,67,73,76]. The cohorts of 6 of these studies consisted of stage IIIA and IIIB patients with a relatively comparable distribution. The other 2 studies consisted mainly or only of stage IIIA patients making them less comparable. One of these 2 was estimated to have a low relevance for utilization of both clinical and pathological staging [57,62]. The patients in the 6 other studies received radiotherapy and/or surgery, and had a relatively comparable distribution of histological subtypes to each other and the other inclusions. A further complication for the comparison was the aforementioned use of different versions of the TNM-staging system. However for this comparison no well-defined subgroup for analysis could be performed, due to the heterogeneity of the studies. For the clinical T1 vs T2 comparison (Fig. 2C), four studies were available for meta-analysis [36,62,73,76]. There was no indication of heterogeneity between the studies (I^2 = 0.00%, τ = 0.00).

The pooled HR was 1.18 (95% CI 0.98–1.41, p = 0.07). For the comparisons T1 vs T3 and T1 vs T4, three studies were available (Fig. 2D, E) [36,62,76]. There was moderate heterogeneity in the T1 vs T3 studies (I^2 = 49.4%, τ = 0.24), the pooled HR was 1.30 (95% CI 0.89–1.91, p = 0.18). For the T1 vs T4 comparison there was also moderate heterogeneity (I^2 = 31.8%, τ = 0.22), the pooled HR was 1.02 (95% CI 0.71–1.46, p = 0.93). These results provide additional evidence that T-stage is not prognostic for OS of stage III NSCLC patients.

The effect of clinical T3-stage on OS in comparison to T1- and T2-stage was measured in 3 studies [37,40,43]. These inclusions consisted solely of stage IIIA patients submitted to surgery. While the study populations were comparable to each other, this reduced their comparability to the overall stage III NSCLC population. However, as 2 of these did not explicitly state the utilized TNM-staging system their comparison was complicated. Significance was reported in univariate analysis of 1 publication [43], while the other 2 reported insignificance [40]. Consequently, included data indicates this comparison is not significant for stage IIIA patients receiving surgery.

Finally, 5 studies with alternative or unspecified comparisons reported T-stage to not be statistically significant for OS in univariate and 1 out of 2 multivariable analyses [46,50,75,79,84]. Four of these reported surgery as a possible treatment modality, and, as 2 consisted only of stage IIIA patients, their relevance was lower.

Table 4. Summary of findings T-stage and lymph node volume.

T-stage and lymph node volume related prognostic factors	Study citation (First author, year (PMID))	Description prognostic factor groups	Number of patients	OS	Univariate analysis (estimate (95% CI))	Multivariable analysis (estimate (95% CI)	Factors corrected for in multivariable analysis

T-stage	Huo X, 2017 (29441096)	T1 + T2 vs T3 + T4	27/155	1y survival rate (%): 96.30/81.10;
				3y survival rate (%): 38.23/23.53;
				5y survival rate (%): 19.66/7.98
					p = 0.037		

	Yilmaz U, 2018 (29559214)	T1 + T2 vs T3 + T4	17/62		HR: 1.273 (0.645–2.513) p = 0.486	HR: 1.565 (0.765–3.201) p = 0.220	PS, Weight loss, N2 disease, N3 disease, Gender, SUVmax
	Koo TR, 2014 (25498887)	T1 + T2 vs T3 + T4	79/78	3y survival rate (%): 43.8/ 28.7	p = 0.106		
	Fried DV, 2016 (26176655)	T1 + T2 vs T3 + T4	97/98			HR: 0.820 p = 0.31	Overall stage, Induction chemotherapy, Age, Gender, GTV, KPS, Co-occurrence matrix energy, Solidity
	Hyun SH, 2014 (23948859)	T1 + T2 vs T3 + T4	125/69				
	Hyun SH, 2014 (23948859)	T1 + T2 vs T3 + T4	125/69		HR: 1.254 (0.829–1.898) p = 0.283	HR: 1.297 (0.702–2.397) p = 0.406	Overall stage, N-stage, ECOG PS, Neoadjuvant chemoradiotherapy, Type of surgery, Chemotherapy, Radiotherapy, SUVmax, MTV
		T1 + T2 vs T3 + T4	125/69			HR: 1.479 (0.825–2.651) p = 0.188	Overall stage, N-stage, ECOG PS, Neoadjuvant chemoradiotherapy, Type of surgery, Chemotherapy, Radiotherapy, SUVmax, TLG
	Wald P, 2017 (28843360)	T1 + T2 vs T3 + T4	22/30		HR: 1.29 (0.54–3.08) p = 0.571		
	Jie Y, 2017 (NA)	T1 + T2 vs T3 + T4	25/53		HR: 1.069 (0.620–1.844) p = 0.810	HR: 1.176 (0.666–2.075) p = 0.577	N-stage, CTV, AUC CSH, SUVmax, MTV, TLG
	Saga T, 2015 (NA)	T0 + T1 + T2 vs T3 + T4	x		HR: 2.39 (0.14–39.71) p = 0.543		
	Dong X, 2016 (27322376)	T1 + T2 vs T3 + T4	25/33		HR: 1.625 (0.282–2.173) p = 0.267		
	Horinouchi H, 2012 (23004347)	T1 + T2 vs T3 + T4	56/54			HR: 0.91 (0.53–1.61) p = 0.77	Age, Gender, Weight loss, Histology, N-stage, Overall stage
	Kanzaki H, 2016 (27125214)	T0 + T1 + T2 vs T3 + T4	59/52		HR: 1.55 (1.00–2.41) p = 0.048		

	Hayakawa K, 1996 (8765179)	T1 + T2 vs T3 vs T4	40/58/43	Median (months): 14/13/10;
				2y survival rate (%): 28/21/23;
				5y survival rate (%): 13/10/9
					p = 0.59		

		T1 + T2 vs T3 + T4	40/101			HR: 1.30 (1.04–1.61) p = 0.021	Age, Gender, PS, Histology, Tumor size, N-stage, Location, Total dose, Field size
	Kim E, 2019 (30266585)	T1 + T2 vs T3 + T4	81/49	3y survival rate (%): 51.8/44.1	p = 0.238		
	Kim DY, 2019 (31591865)	T1 + T2 vs T3 + T4	35/47		HR: 0.86 (0.53–1.40) p = 0.538		

Table 4. Continued.

T-stage and lymph node volume related prognostic factors	Study citation (First author, year (PMID))	Description prognostic factor groups	Number of patients	OS	Univariate analysis (estimate (95% CI))	Multivariable analysis (estimate (95% CI)	Factors corrected for in multivariable analysis
	Hyun SH 2015 (26295651)	T1 + T2 vs T3	132/29		HR: 2.50 (1.31–4.78) p = 0.005		

	Li J, 2009 (NA)	T1 + T2 vs T3	53/38	Median (months): 32/27;
				1y survival rate (%): 88.9/86.7;
				3y survival rate (%): 44.4/ 26.7;
				5y survival rate (%): 29.2/22.0			

	Betticher DC, 2006 (16622435)	T1 + T2 vs T3	50/25	Median (months): 27.6/57.1	p = 0.12		

	Sibley GS, 1995 (7493826)	T1 vs T2 vs T3 vs T4	6/7/11/13	Median (months): unkown/12.1/20.4/19.5;
				1y survival rate (%): 67/71/73/84;
				2y survival rate (%): 67/29/45/32
					p = 0.52		

	Basaki K, 2006 (16226400)	T1 vs T2 vs T3 vs T4	4/18/23/26	Median (months): 11/12/13/14; 2y survival rate (%): 28/23/20/23	Not significant	HR: 0.73 (0.51–1.04) p = 0.08	Primary tumor volume/Total tumor volume, N-stage, PS, 2 Gy equivalent dose, Age, Chemotherapy, Histology, Nstage
	Agrawal V, 2017 (28426673)	T1 vs T2	18/32		HR: 0.77 (0.35–1.71) p = 0.53		
		T1 vs T3 + T4	18/23		HR: 0.40 (0.14–1.10) p = 0.08		
	Zhou R, 2018 (NA)	T0 + T1 vs T2	92/175		HR: 1.09 (0.775–1.541) p = 0.614		
		T0 + T1 vs T3	92/85		HR: 1.177 (0.785–1.764) p = 0.431		
		T0 + T1 vs T4	92/125		HR: 1.217 (0.848–1.747) p = 0.286		
	Oberije C, 2015 (25936599)	T0 + T1 vs T2			HR: 1.11 (0.81–1.52) p = 0.3135		
		T0 + T1 vs T3			HR: 1.8 (0.92–2.07) p = 0.3135		
		T0 + T1 vs T4			HR: 1.06 (0.76–1.50) p = 0.3135		
	Tao X, 2019 (31179087)	T1 vs T2	271/239		HR: 1.41 (1.04–1.90) p = 0.0265	HR: 1.22 (0.89–1.67) p = 0.2181	Age, Gender, Smoking history, Tumor location, Treatment approach, N-stage
		T1 vs T3	271/58		HR: 0.90 (0.48–1.69) p = 0.7459	HR: 0.83 (0.43–1.58) p = 0.5667	Age, Gender, Smoking history, Tumor location, Treatment approach, N-stage

		T1 vs T4	271/35		HR: 0.57 (0.27–1.31) p = 0.1872	HR: 0.51 (0.22–1.21) p = 0.1290	Age, Gender, Smoking history, Tumor location, Treatment approach, N-stage
	Maniwa T, 2018 (30746228)	T1 vs T2 + T3 + T4	30/64	5y survival rate (%): 54.0/41.8	p = 0.39		
	Hwang IG, 2008 (18623378)	T1 vs T2 + T3	12/56	Median (months): 42.6/41.7	p = 0.687		
	Broderick SR, 2016 (26410162)	NR				p = 0.135	Age, Gender, Ethnicity, Income, Treatment, Charlson/Deyo score, Tumor size, Neoadjuvant therapy, Right pneumonectomy
	Shumway, 2011 (21676484)	NR			Not significant		
	Akcam TI, 2015 (NA)	NR		Median (months): 14	p = 0.053		
Nodal volume	Alexander BM, 2011 (20605346)	Continuous: by 10 cm^3 increase (All patients)	107	Median (months): 23	HR: 1.09 p < 0.01	HR: 1.09 (1.05–1.13) p < 0.01	Overall stage, Surgery, Radiation dose
		Continuous: by 10 cm^3 increase (Chemoradiation patients)	76	Median (months): 15	HR: 1.07 p < 0.01	HR: 1.09 (1.05–1.14) p < 0.01	Gender, Tumor volume
	Basaki K, 2006 (16226400)	<15 cm^3 vs >15 cm^3	35/36	Median: 14/12 months; 2 year survival rate (%): 24/19	Not significant	HR: 1.06 (0.99–1.14) p = 0.10	Primary tumor volume, PS, 2 Gy equivalent dose, Chemotherapy, Age, Histology, T-stage, N-stage
	Lee VHF, 2016 (24710123)	Continuous	43	Median (months): 37.8	p = 0.402		
	Oberije C, 2015 (25936599)	Continuous (mL)	548		HR: 1.16 (1.14–1.18) p = 0.0008		

	Dieleman EMT, 2018 (30055239)	Continuous	154	Median (months): 36.1;
				1y survival rate (%): 79;
				2y survival rate (%): 61;
				3y survival rate (%): 52;
				5y survival rate (%): 40
					HR: 1.004 (1.00–1.008)
					p = 0.033
						HR: 1.007 (1.0–1.012)
						p = 0.047
							Gender, Age, Radiation technique

Nodal diameter	Betticher DC, 2006 (16622435)	≤1cm vs >1 cm	14/61	Median (months): 32.5/29.9	p = 0.47		

Outcomes concerning T-stage and nodal volume in univariate and multivariable analysis of the included studies. Estimates are reported with 95% confidence interval and p-value when available. Used statistical models and tests are: Cox proportional hazard model, log-rank test (LR). Abbreviations: AUC-CSH: Area under the curve of cumulative SUV histograms, CI: Confidence interval, ECOG: Eastern cooperative oncology group, GTV: Gross tumor volume, Gy: Gray, HR: Hazard ratio, (K)PS: (Karnofsky) performance status, MTV: Metabolic tumor volume, NA: Not applicable, NR: Not reported, OS: Overall survival, PMID: PubMed identification number, PS: Performance score, SUV: Standardized uptake volume, TLG: Total lesion glycolysis, Y: year.

Two prognostic factors specifically concern the involved lymph nodes. Nodal volume was measured in 5 studies (Table 4) [18,36,66,80,81]. Regardless of the presence of patients receiving surgery, nodal volume was found to be significant in univariate and multivariable analysis of 3 studies [36,66,81], but not in univariate and multivariable analysis of the other 2 [18,80]. The studies were comparable to the subgroup analysis for patients who received exclusively chemoradiation, in distribution of stage IIIA/ IIIB and histological subtypes. Considering all this, total lymph node volume is likely to be a prognostic factor for stage III patients.

Dichotomized nodal diameter was not found to be significant in univariate analysis of a single study [37]. This study consisted of only stage IIIA patients treated with surgery, and was therefore not representative for the standard stage III population. As this only concerns a single study, no definite conclusions can be drawn.

Prognostic factors that could not be classified as size, nodal, or texture-related, were classified as other CT-related prognostic factors (Table 2). These included 8 unique factors: Atelectasis/ Obstructive pneumonitis [20,67], Location [18,34,42,48,49,52, 62,63,71,72,75,78], Cavitary wall thickness [17], Cavitation [33,44,51], Interstitial lung abnormalities [55], Great vessel invasion [82], Pit fall sign [14], and Pleural effusion [39].

Atelectasis was studied in 2 inclusions, in 1 as a dichotomous factor [20] and in the other as a discrete variable with more than 2 levels [67]. Atelectasis did not yield significance in univariate analysis as a discrete factor [67], but did as a dichotomous factor [20]. Both the studies consisted of stage IIIA and IIIB patients receiving chemoradiation. The representativeness of the publication that considered atelectasis as a dichotomous factor for the entire stage III NSCLC population cannot be fully assessed, as it did not report the distribution of histological subtype [20]. Additionally, the relevance of publication that considered atelectasis as a discrete factor was decreased, as it made use of an older version of the TNM staging system. Due to these issues in the 2 publications, no concrete conclusion can be made.

Twelve studies reported data on the effect of tumor location on OS in several discrete ways [18,34,42,48,49,52,62,63,71,72,75,78]. Four inclusions compared presence in the right and left lung [18,49,63,72], another 2 between central and peripheral location [62,71]. For both comparisons no significance was reported in univariate analysis. However 1 study for each of the 2 respective comparisons was estimated to have a low relevance on behalf of consisting of clinically as well as pathologically staged III patients [62,63]. As a consequence, considering most of the other studies seemed to be representative for the overall stage III NSCLC population [18,71,72], the inclusions give little reason for future research of left/right location. The final comparison was between pulmonary lobes, for which a significant correlation was found in 2 out of 5 univariate [42,48] and 2 out of 3 multivariable analyses [34,48]. It should be noted, however, that in 3 studies which found no significance and 1 which found significance, patients were treated with surgery, decreasing their comparability to the other study cohorts [42,52,78]. Concluding, considering the heterogeneity of the inclusions data, regarding both central/peripheral location and tumor location by lobes remains inconclusive.

Two prognostic factors concern cavitation: appearance of a region with lower density within the tumor mass. Cavitation itself was studied in 3 publications, in which it was reported to be significant in 1 out of 3 univariate analyses and in multivariable analysis [33,44,51]. It should however be noted that 1 study, in which no significance was found, consisted only of IIIA patients treated with surgery. Cavitary wall thickness was reported not to be a significant prognostic factor in a subgroup analysis of a single study for stage III patients treated with surgery [17]. However, this cohort was not representative for the overall stage III population, consisting exclusively of adenocarcinoma patients, and because tumor cavitation is present in less than 25% of lung cancer cases [86]. However, due to the relatively limited data no definite conclusions can be drawn about factors concerning cavitation.

The last four CT-related prognostic factors were measured in single studies. Both interstitial lung abnormalities and great vessel invasion were reported to be not significant as prognostic factors in a stage III NSCLC cohort [55,82]. Pit fall sign, studied in subgroup analyses for stage III NSCLC patients treated with surgery, was not found to be significant. However, these results were based on only 16 stage III patients and should be verified in a larger stage III cohort [14]. The effect of pleural effusion, analyzed in a stage IIIA and IIIB specific manner, was reported to be significant in univariate analysis in stage IIIA patients, and in both univariate and multivariable analysis in stage IIIB patients [39].

Discussion and conclusion

In this systematic review and meta-analysis, 26 unique CTrelated prognostic factors were identified for OS in 65 studies comprising 144,513 stage III NSCLC patients. Inclusions indicated Tstage is unlikely to be prognostic for OS of stage III NSCLC patients treated with chemoradiation, as it was found to be insignificant in the majority of analyses [15,18,21,31,36,37,40,43,46,48,50,53,54,56–59,61,62,65,67,71–73,75,76,79,84,85]. Although population characteristics of publications concerning size-related prognostic factors were heterogeneous, there was an indication that GTV, tumor diameter, and nodal volume are prognostic for OS of stage III patients receiving chemoradiation [11,13,15,18,21,32,34–36,41,51,58,64–68,70,71,74,77,80,81,83], but that this may not be the case for tumor volume and diameter in cohorts containing NSCLC patients receiving surgery [16,38,66,68]. This could potentially be explained by the aim of surgery to remove the tumor and involved lymph nodes, which could conceivably undermine size-related prognostic effects [87]. While tumor diameter and volume are related, it is notable that we could not draw any conclusions regarding tumor volume for stage III patients receiving chemoradiation. This was mainly caused by the heterogeneity of the included data, which also hampered the analysis of other factors including atelectasis and location (by pulmonary lobe). The exact extent of heterogeneity in the data is discussed below [16,18,32,34,36–38,45,48–52,58,60,66,68,69,78,80,81]. Furthermore, T-stage, which is partially determined by tumor size as proposed by the international association for the study of lung cancer [88,89], did not seem to hold prognostic value within NSCLC cohorts consisting solely of stage III patients, while GTV and tumor diameter did. A potential explanation is that in cohorts restricted to stage III patients Nstage is dominant in OS of patients with smaller tumors. Additionally restricting the analysis to stage III patients may reduce the variation in T-stage between patients to greater extent than it reduces variation in tumor size, as the T-stage directly influences overall stage. Similarly, a decrease in T-stage necessarily entails an increase in N-stage for stage III patients, lowering the relevance of univariate prognostic models of these factors.

The 2 included studies concerning radiomic features suggest several features (including entropy, skewness, mean HU, largest axial slice average, largest axial slice uniformity, HU kurtosis, HU infomc1, HU standard deviation, and HU sosvariance) have potential prognostic value for stage III NSCLC patients receiving chemoradiation. However, considering this concerned only 2 studies and the vulnerability of radiomic features to difficulties in validation [90,91], we feel this group of prognostic factors warrant separate review. These factors should be validated in a larger cohort [11,12]. Finally, of the other CT-related prognostic factors, location (right/left) is not likely to be a prognostic factor [18,71,72]. Pleural effusion did, however, seem to be a prognostic factor in a single study [39]. No concrete conclusions could be drawn concerning atelectasis, cavitation, and location (by pulmonary lobes, central/peripheral), as evidence was too heterogeneous [18,20,33,42,44,51,67], or for cavitary wall thickness and pit fall sign, as the stage III subgroup of their studies was not representative for the standard stage III NSCLC population [14,17]. More research is warranted to validate these results.

This study presents an overview of prognostic factors for OS of stage III NSCLC patients. Several potential prognostic factors were identified, which could be used to direct future research. Several factors hamper the strength of the conclusions that can be drawn from this systematic review. In 32 studies the utilized staging method (clinical/pathological) was not specified [11,13,15,17,18,20,32,34,35,37,64–80,85]. Three inclusions even compared patients with pathological and clinical stage III [61–63].We recommend that future studies into prognostic factors are reported according to the TRIPOD reporting guidelines to increase their scientific value and facilitate the use of their results in meta-analysis [24]. Additionally clinical staging is preferred to pathological staging, because, even though in theory pathological stage correlates better to prognosis, ultimately only clinical stage is available for treatment decisions [9,92,93].

Another limitation was that CT-related prognostic factors were not often the primary focus of the included articles. This may have led to relevant articles not being retrieved with the utilized search terms.

We were unable to estimate the risk of publication bias from the provided data due to the low number of studies per prognostic factor. As virtually all studies reported the results on multiple prognostic factors instead of just one, it is less likely that a nonsignificant result for one of the prognostic factors would have reduced the probability of publication. However, for continuous prognostic factors or prognostic factors with multiple categories, there are several ways to include this variable in the analysis. The way a variable was entered in the analysis (e.g. dichotomized GTV or choosing groups of T-stage for comparison) could be driven by the data and reasons behind these choices were hardly ever reported. This increases the risk of false positive findings.

Additionally, inclusions were found to be heterogeneous in distribution of histological subtypes, stage IIIA/IIIB, and treatment modalities. This limited the analysis of several prognostic factors including atelectasis [20,67], and location (by pulmonary lobes) [18,42]. It should also be noted that surgery was reported to be a treatment option in 25 of the 65 inclusions [14,16,17,33,37–40,4 2–44,46,49,50,52,57,61–63,66,68,73,75,78,79]. Considering surgery might influence the relevance of size-related prognostic factors [92,93], these studies may not be comparable to stage III cohorts receiving chemoradiation alone. Finally, OS was measured from distinct time points [12–14,18,21,31–33,35,36,38,40–45,47,48,51,52,59,60,64–66,71–74,77,85]: where some used OS measured from the first day of chemoradiation treatment onwards [12,13,18,32,42–44,55,57,59,60,63–65,71–73,81–85], others measured OS from time of diagnosis [16,20,34,37,39,46,49,50,53,56, 61,67–69,78,79]. This complicates comparisons between study cohorts.

Notably, only 6 studies included weight loss in multivariable analysis, even though it is a prognostic factor recognized by guidelines [39,45,51,53,59,74]. Moreover, performance status was included in only 12 of 65 publications [11,18,34,39,45,48,59,61,65,70,74]. The value of new prognostic markers should be evaluated in light of existing ones. It is recommended for future research to explicitly include comparisons with the established prognostic markers weight loss and performance status.

Considering these heterogeneities between the included studies, which hampered our ability to come to strong conclusions concerning both the significance and clinical relevance of the aforementioned prognostic factors, including tumor volume, we suggest future studies report the employed staging system (clinical or pathological, and TNM version), received treatments, presence and handling of missing data, effects sizes, and measures of uncertainty such as confidence intervals. Additionally we advise studies concerning radiomic features to carefully describe the methods used to obtain the results, for reproducibility and future data analysis, specifically in the ways suggested by Zwanenburg et al. (2020) [90] and Welch et al. (2019) [91]. Finally, future studies should compare the measured prognostic factor with those recognized by the clinical guidelines (weight loss and performance status) and validated prognostic factors from other studies.

In conclusion, Gross Tumor Volume, tumor diameter, nodal volume, and pleural effusion are likely to be prognostic factors for OS of stage III patients treated with chemoradiation. Several radiomic features have potential prognostic value. Additionally, the combined evidence strongly indicates that T-stage and location (right/ left) are not prognostic for OS within the group of stage III NSCLC patients. Finally, the included evidence concerning tumor volume, atelectasis, location (by pulmonary lobes, central/peripheral), pit fall sign, and cavitation remains inconclusive. Regarding these prognostic factors, more research is needed before firm conclusions can be made and clinically relevant prognostic factors could be used to improve treatment decisions. To improve the evaluation of evidence, future studies should both carefully report the employed staging system, received treatments, effects sizes and measures of uncertainty, and contrast the measured prognostic factor with guideline recognized prognostic factors in addition to those from earlier studies, as presented in this systematic review.

Acknowledgements

Thanks are due to the UMCU librarians, specifically Felix Weijdema, for help with the search terms.

Appendix A. Supplementary data

Supplementary data to this article can be found online at *https://doi.org/10.1016/j.radonc.2020.07.030*.

References

[1] Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin 2015;65:87–108. https://doi.org/10.3322/caac.21262.

[2] Stewart BW, Kleihues P. World cancer report. Lyon: IACR Press; 2003.

[3] Ervik M, Lam F, Ferlay J, Mery L, Soerjomataram I, Bray F. Cancer today. Cancer Today Lyon, Fr Int Agency Res Cancer 2016. http://gco.iarc.fr/today (accessed June 9, 2018).

[4] Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, et al. Cancer incidence and mortality worldwide: IACR CancerBase No. 11. GLOBOCAN 2012. http://globocan.iarc.fr (accessed June 9, 2018).

[5] Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: Epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc 2008;83:584–94. https://doi.org/10.1080/10810730902873927. Testing.

[6] de PI, Oliveira, Pereira CA de C, Belasco AGS, Bettencourt AR de C. Comparison of the quality of life among persons with lung cancer, before and after the chemotherapy treatment. Rev Lat Am Enfermagem 2013;21:787–94. https://doi.org/10.1590/S0104-11692013000300019.

[7] Aupérin A, Le Péchoux C, Rolland E, Curran WJ, Furuse K, Fournel P, et al. Metaanalysis of concomitant versus sequential radiochemotherapy in locally advanced non-small-cell lung cancer. J Clin Oncol 2010;28:2181–90. https://doi.org/10.1200/JCO.2009.26.2543.

[8] National Comprehensive Cancer Network. NCCN clinical practice guidelines in oncology non-small cell lung cancer. Version 2. 2019.

[9] Beadsmoore CJ, Screaton NJ. Classification, staging and prognosis of lung cancer. Eur J Radiol 2003;45:8–17. https://doi.org/10.1016/S0720-048X(02) 00287-5.

[10] Gridelli C, Rossi A, Carbone DP, Guarize J, Karachaliou N, Mok T, et al. Nonsmall-cell lung cancer. Nat Rev Dis Prim 2015;1:1–16. https://doi.org/10.1038/nrdp.2015.9.

[11] Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, et al. Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. Int J Radiat Oncol Biol Phys 2014;90:834–42. https://doi.org/10.1016/j.ijrobp.2014.07.020.

[12] Ahn SY, Park CM, Park SJ, Kim HJ, Song C, Lee SM, et al. Prognostic value of computed tomography texture features in non-small cell lung cancers treated with definitive concomitant chemoradiotherapy. Invest Radiol 2015;50:719–25. https://doi.org/10.1097/RLI.0000000000000174.

[13] Elsayad K, Samhouri L, Scobioala S, Haverkamp U, Eich HT. Is tumor volume reduction during radiotherapy prognostic relevant in patients with stage III non-small cell lung cancer? J Cancer Res Clin Oncol 2018;144:1165–71. https://doi.org/10.1007/s00432-018-2640-6.

[14] Li M, Ito H, Wada H, Tanaka F. Pit-fall sign on computed tomography predicts pleural involvement and poor prognosis in non-small cell lung cancer. Lung Cancer 2004;46:349–55. https://doi.org/10.1016/j.lungcan.2004.05.017.

[15] Koo TR, Moon SH, Lim YJ, Kim JY, Kim Y, Kim TH, et al. The effect of tumor volume and its change on survival in stage III non-small cell lung cancer treated with definitive concurrent chemoradiotherapy. Radiat Oncol 2014;9:283–90. https://doi.org/10.1186/s13014-014-0283-6.

[16] Soussan M, Chouahnia K, Maisonobe JA, Boubaya M, Eder V, Morère JF, et al. Prognostic implications of volume-based measurements on FDG PET/CT in stage III non-small-cell lung cancer after induction chemotherapy. Eur J Nucl Med Mol Imaging 2013;40:668–76. https://doi.org/10.1007/s00259-012-2321-7.

[17] Watanabe Y, Kusumoto M, Yoshida A, Shiraishi K, Suzuki K, Watanabe S, et al. Cavity wall thickness in solitary cavitary lung adenocarcinomas is a prognostic indicator. Ann Thorac Surg 2016;102:1863–71. https://doi.org/10.1016/j.athoracsur.2016.03.121.

[18] Basaki K, Abe Y, Aoki M, Kondo H, Hatayama Y, Nakaji S. Prognostic factors for survival in stage III non-small-cell lung cancer treated with definitive radiation therapy: Impact of tumor volume. Int J Radiat Oncol Biol Phys 2006;64:449–54. https://doi.org/10.1016/j.ijrobp.2005.07.967.

[19] Kozak MM, Murphy JD, Schipper ML, Donington JS, Zhou L, Whyte RI, et al. Tumor volume as a potential imaging-based risk-stratification factor in trimodality therapy for locally advanced non-small cell lung cancer. J Thorac Oncol 2011;6:920–6. https://doi.org/10.1097/JTO.0b013e31821517db.

[20 Bulbul Y, Eris B, Orem A, Gulsoy A, Oztuna F, Ozlu T, et al. Pulmonary atelectasis and survival in advanced non-small cell lung carcinoma. Ups J Med Sci 2010;115:176–80. https://doi.org/10.3109/03009731003695624.

[21] Wald P, Mo X, Barney C, Gunderson D, Haglund AK, Bazan J, et al. Prognostic value of primary tumor volume changes on kV-CBCT during definitive chemoradiotherapy for stage III non–small cell lung cancer. J Thorac Oncol 2017;12:1779–87. https://doi.org/10.1016/j.jtho.2017.08.010.

[22] Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. Proc 2nd ACM SIGHIT Symp Int Heal Informatics – IHI'12 2012:819–24. https://doi.org/10.1145/2110363.2110464.

[23] The Cochrane Collaboration. Cochrane handbook for systematic Reviews of interventions. Version 5. 2011.

[24] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyenberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or disagnosis (TRIPOD): Explanantion and elaboration. Ann Intern Med 2015;162:W1–W73. https://doi.org/10.7326/M14-0698.

[25] Scottish Intercollegiate Guidelines Network. SIGN 50: a guideline developer's handbook. Edinburgh: 2015.

[26] Paule R, Mandel J. Consensus values, regressions, and weighting factors. J Res Natl Inst Stand Technol 1989;94:197–203.

[27] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsisteny in meta-analyses. Br Med J 2003;327:557–60.

[28] Higgins JP, Thompson SG. Controlling the risk of spurious findings from metaregression. Stat Med 2004;23:1663–82.

[29] Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health 2019;22:153–60. https://doi.org/10.1136/ebmental-2019-300117.

[30] Harrer M, Cuijpers P, Furukawa T, Ebert D. dmetar: Companion R package for the guide "doing meta-anaylisis in R". R package version 0.0.9000 2019.

[31] Saga T, Inubushi M, Koizumi M, Yoshikawa K, Zhang M-R, Tanimoto K, et al. Prognostic value of 18F-fluoroazomycin arabinoside PET/CT in patients with advanced non-small-cell lung cancer. Cancer Sci 2015;106:1554–60. https://doi.org/10.1111/cas.12771.

[32] Crvenkova S. Prognostic factors and survival in non-small cell lung cancer patients treated with chemoradiotherapy. Open Access Maced J Med Sci 2015;3:75–9. https://doi.org/10.3889/oamjms.2015.003.

[33] Chaft JE, Rusch V, Ginsberg MS, Paik PK, Finley DJ, Kris MG, et al. Phase II trial of neoadjuvant bevacizumab plus chemotherapy and adjuvant bevacizumab in patients with resectable nonsquamous non–small-cell lung cancers. JTO Acquis 2013;8:1084–90. https://doi.org/10.1097/JTO.0b013e31829923ec.

[34] Chang JY, Verma V, Li M, Zhang W, Ritsuko K, Lu C, et al. Proton beam radiotherapy and concurrent chemotherapy for unresectable stage III nonsmall cell lung cancer: final results of a phase 2 study. JAMA. Oncol 2017;3. https://doi.org/10.1001/jamaoncol.2017.2032.

[35] Nguyen Q, Ly NB, Komaki R, Levy LB, Daniel R, Chang JY, et al. Long-term outcomes after proton therapy, with concurrent chemotherapy, for stage II-III inoperable non-small cell lung cancer. Radiother Oncol 2015;115:367–72. https://doi.org/10.1016/j.radonc.2015.05.014.

[36] Oberije C, De Ruysscher D, Houben R, van de Heuvel M, Uyterlinde W, Deasy JO, et al. A validated prediction model for overall survival from stage III nonsmall cell lung cancer: toward survival prediction for individual patients. Int J Radiat Oncol Biol Phys 2015;92:935–44. https://doi.org/10.1002/cncr.27633. Percutaneous.

[37] Betticher DC, Hsu Schmitz SF, Tötsch M, Hansen E, Joss C, Von Briel C, et al. Prognostic factors affecting long-term outcomes in patients with resected stage IIIA pN2 non-small-cell lung cancer: 5-Year follow-up of a phase II study. Br J Cancer 2006;94:1099–106. https://doi.org/10.1038/sj.bjc.6603075.

[38] Lee HY, Lee KS, Park J, Han J, Kim BT, Kwon OJ, et al. Baseline SUVmax at PETCT in stage IIIA non-small-cell lung cancer patients undergoing surgery after neoadjuvant therapy: Prognostic implication focused on histopathologic subtypes. Acad Radiol 2012;19:440–5. https://doi.org/10.1016/j.acra.2011.12.010.

[39] Ryu J-S, Ryu HJ, Lee S-N, Memon A, Lee S-K, Nam H-S, et al. Prognostic impact of minimal pleural effusion in non-small-cell lung cancer. J Clin Oncol 2014;32:960–7. https://doi.org/10.1200/JCO.2013.50.5453.

[40] Hyun SH, Ahn HK, Ahn MJ, Ahn YC, Kim J, Shim YM, et al. Volume-based assessment with 18F-FDG PET/CT improves outcome prediction for patients with stage IIIA-N2 non-small cell lung cancer. Am J Roentgenol 2015;205:623–8. https://doi.org/10.2214/AJR.14.13847.

[41] Xiang Z-L, Erasmus J, Komaki R, Cox JD, Chang JY. FDG uptake correlates with recurrence and survival after treatment of unresectable stage III non-small cell lung cancer with high-dose proton therapy and chemotherapy. Radiat Oncol 2012;7:144–51. https://doi.org/10.1186/1748-717X-7-144.

[42] Shien K, Toyooka S, Soh J, Hotta K, Katsui K, Oto T, et al. Lower lobe origin is a poor prognostic factor in locally advanced non-small-cell lung cancer patients treated with induction chemoradiotherapy. Mol Clin Oncol 2015;3:706–12. https://doi.org/10.3892/mco.2015.509.

[43] Li J, Dai C-H, Shi S-B, Chen P, Yu L-C, Wu J-R. Prognostic factors and long term results of neo adjuvant therapy followed by surgery in stage IIIA N2 non-small cell lung cancer patients. Ann Thorac Med 2009;4:201–7. https://doi.org/10.4103/1817-1737.56010.

[44] Phernambucq ECJ, Hartemink KJ, Smit EF, Paul MA, Postmus PE, Comans EFI, et al. Tumor cavitation in patients with stage III non-small-cell lung cancer undergoing concurrent chemoradiotherapy. J Thorac Oncol 2012;7:1271–5. https://doi.org/10.1097/JTO.0b013e3182582912.

[45] Naito Y, Kubota K, Nihei K, Fuji T, Yoh K, Niho S, et al. Concurrent chemoradiotherapy with cisplatin and vinorelbine for stage III non-small cell lung cancer. J Thorac Oncol 2008;3:617–22. https://doi.org/10.1097/JTO.0b013e3181753b38.

[46] Shumway D, Corbin K, Salgia R, Hoffman P, Villaflor V, Malik RM, et al. Lung cancer pathologic response rates following definitive dose image-guided chemoradiotherapy and resection for locally advanced non-small cell lung cancer. Lung Cancer 2011;74:446–50. https://doi.org/10.1016/j.lungcan.2011.05.003.

[47] Park YJ, Yoon WS, Lee JA, Lee NK, Lee S, Yang DS, et al. Concurrent chemoradiotherapy in locally advanced non-small cell lung cancer: a retrospective analysis of the correlation between radiotherapy-related factors and tumor response. Int J Radiat Res 2015;13:205–12.

[48] Hayakawa K, Mitsuhashi N, Saito Y, Furuta M, Nakayama Y, Katano S, et al. Impact of tumor extent and location on treatment outcome in patients with stage III non-small cell lung cancer treated with radiation therapy. Jpn J Clin Oncol 1996;26:221–8. https://doi.org/10.1093/oxfordjournals.jjco.a023218.

[49] Pang Z, Yang Y, Ding N, Huang C, Zhang T, Ni Y, et al. Optimal managements of stage IIIA (N2) non-small cell lung cancer patients: a population-based survival analysis. J Thorac Dis 2017;9:4046–56. https://doi.org/10.21037/jtd.2017.10.47.

[50] Broderick SR, Patel AP, Crabtree TD, Bell JM, Morgansztern D, Robinson CG, et al. Pneumonectomy for clinical stage IIIA non-small cell lung cancer: The impact of neoadjuvant therapy. Ann Thorac Surg 2016;101:451–8. https://doi.org/10.1016/j.athoracsur.2015.07.022.

[51] Topkan E, Selek U, Ozdemir Y, Yildirim BA, Guler OC, Ciner F, et al. Incidence and impact of pretreatment tumor cavitation on survival outcomes of stage III squamous cell lung cancer patients treated with radical concurrent chemoradiation therapy. Int J Radiat Oncol Biol Phys 2018;101:1123–32. https://doi.org/10.1016/j.ijrobp.2018.04.053.

[52] Hishida T, Yoshida J, Ohe Y, Aokage K, Ishii G, Nagai K. Surgical outcomes after initial surgery for clinical single-station N2 non-small-cell lung cancer. Jpn J Clin Oncol 2014;44:85–92. https://doi.org/10.1093/jjco/hyt164.

[53] Horinouchi H, Sekine I, Sumi M, Noda K, Goto K, Mori K, et al. Long-term results of concurrent chemoradiotherapy using cisplatin and vinorelbine for stage III non-small-cell lung cancer. Cancer Sci 2013;104:93–7. https://doi.org/10.1111/cas.12028.

[54] Kanzaki H, Kataoka M, Nishikawa A, Uwatsu K, Nagasaki K, Nishijima N, et al. Impact of early tumor reduction on outcome differs by histological subtype in stage III non-small-cell lung cancer treated with definitive radiotherapy. Int J Clin Oncol 2016;21:853–61. https://doi.org/10.1007/s10147-016-0982-0.

[55] Higo H, Kubo T, Makimoto S, Makimoto G, Ihara H, Masaoka Y, et al. Chemoradiotherapy for locally advanced lung cancer patients with interstitial lung abnormalities. Jpn J Clin Oncol 2019;49:458–64. https://doi.org/10.1093/jjco/hyz016.

[56] Kim E, Wu H, Keam B, Kim TM, Kim D, Paeng JC, et al. Significance of 18 F-FDG PET parameters according to histologic subtype in the treatment outcome of stage III non-small-cell lung cancer undergoing definitive concurrent chemoradiotherapy. Clin Lung Cancer 2019;20:9–23.

[57] Maniwa T, Shintani Y, Okami J, Kadota Y, Takeuchi Y, Takami K, et al. Upfront surgery in patients with clinical skip N2 lung cancer based on results of modern radiological examinations. J Thorac Dis 2018;10:6828–37. https://doi.org/10.21037/jtd.2018.10.115.

[58] Huo X, Huo B, Wang H, Wang L, Cao Q, Zheng G, et al. Implantation of computed tomography-guided Iodine-125 seeds in combination with chemotherapy for the treatment of stage III non-small cell lung cancer. J Contemp Brachyther 2017;9:527–34. https://doi.org/10.5114/jcb.2017.72605.

[59] Yılmaz U, Batum Ö, Koparal H, Özbilek E, Kıraklı E. Prognostic value of primary tumor SUVmax on pre-treatment 18F-FDG PET/CT imaging in patients with stage III non-small cell lung cancer. Rev Española Med Nucl e Imagen Mol (English Ed 2018. https://doi.org/10.1016/j.remnie.2017.12.001.

[60] Firat S, Byhardt RW, Gore E. Comorbidity and Karnofksy performance score are independent prognostic factors in stage III non-small-cell lung cancer: an institutional analysis of patients treated on four RTOG studies. Int J Radiat Oncol Biol Phys 2002;54:357–64.

[61] Hyun SH, Ahn HK, Kim H, Ahn M-J, Park K, Ahn YC, et al. Volume-based assessment by 18F-FDG PET/CT predicts survival in patients with stage III nonsmall-cell lung cancer. Eur J Nucl Med Mol Imaging 2014;41:50–8. https://doi.org/10.1007/s00259-013-2530-8.

[62] Tao X, Yuan C, Zheng D, Ye T, Yunjian P, Zhang Y, et al. Outcomes comparison between neoadjuvant chemotherapy and adjuvant chemotherapy in stage IIIA non-small cell lung cancer patients. J Thorac Dis 2019;11:1443–55. https://doi.org/10.21037/jtd.2019.03.42.

[63] Casiraghi M, Guarize J, Sandri A, Maisonneuve P, Brambilla D, Romano R, et al. Pneumonectomy in stage IIIA-N2 NSCLC: should it be considered after neoadjuvant chemotherapy? Clin Lung Cancer 2019;20:97–107.

[64] Gensheimer MF, Hong JC, Chang-Halpenny C, Zhu H, Eclov NCW, To J, et al. Mid-radiotherapy PET/CT for prognostication and detection of early progression in patients with stage III non-small cell lung cancer. Radiother Oncol 2017;125:338–43. https://doi.org/10.1016/j.radonc.2017.08.007.

[65] Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, et al. Stage III nonsmall cell lung cancer: prognostic value of FDG PET quantitative imaging features combined with clinical prognostic factors. Radiology 2016;278:214–22. https://doi.org/10.1148/radiol.2015142920.

[66] Alexander BM, Othus M, Caglar HB, Allen AM. Tumor volume is a prognostic factor in non-small-cell lung cancer treated with chemoradiotherapy. Int J Radiat Oncol Biol Phys 2011;79:1381–7. https://doi.org/10.1016/j.ijrobp.2009.12.060.

[67] Sibley GS, Jacobs R, Chen G, Ph D, Weichselbaijm R, Reese M. The treatment of stage III nonsmall cell lung cancer using high dose conformal radiotherapy. Int J Radiat Oncol Biol Phys 1995;33:1001–7.

[68] William WN, Lin HY, Lee JJ, Lippman SM, Roth JA, Kim ES. Revisiting stage IIIB and IV non-small cell lung cancer: Analysis of the surveillance, epidemiology, and end results data. Chest 2009;136:701–9. https://doi.org/10.1378/chest.08-2968.

[69] Morgensztern D, Waqar S, Subramanian J, Gao F, Trinkaus K, Govindan R. Prognostic significance of tumor size in patients with stage III non-small-cell lung cancer: A surveillance, epidemiology, and end results (SEER) survey from 1998 to 2003. J Thorac Oncol 2012;7:1479–84. https://doi.org/10.1097/JTO.0b013e318267d032.

[70] Wu J, Gensheimer MF, Dong X, Rubin DL, Napel S, Diehn M, et al. Robust intratumor partitioning to identify high-risk subregions in lung cancer: A pilot study. Int J Radiat Oncol Biol Phys 2016;95:1504–12. https://doi.org/10.1016/j.ijrobp.2016.03.018.

[71] Jie Y, Meng X, Gu A, Sun X, Yu J. Metabolic volume parameters based on different thresholds with baseline 18F-FDG PET/CT as prognostic factors for survival in stage III non-small cell lung cancer. Transl Cancer Res 2017;6:732. https://doi.org/10.21037/tcr.2017.06.52.

[72] Dong X, Sun X, Sun L, Maxim PG, Xing L, Huang Y, et al. Early change in metabolic tumor heterogeneity during chemoradiotherapy and its prognostic value for patients with locally advanced non-small cell lung cancer. PLoS ONE 2016;11:1–14. https://doi.org/10.1371/journal.pone.0157836.

[73] Agrawal V, Coroller TP, Hou Y, Lee SW, Romano JL, Baldini EH, et al. Lymph node volume predicts survival but not nodal clearance in Stage IIIA-IIIB NSCLC. PLoS ONE 2017;12. https://doi.org/10.1371/journal.pone.0174268.

[74] Etiz D, Marks LB, Zhou S-M, Bentel GC, Clough R, Hernando ML, et al. Influence of tumor volume on survival in patients irradiated for non-small-cell lung cancer. Int J Radiat Oncol Biol Phys 2002;53:835–46.

[75] Akcam TI, Kaya SO, Akcay O, Samancilar O, Ceylan KC, Sevinc S, et al. Is there a survival difference between single station and multi-station N2 disease in operated non-small cell lung cancer patients? Cancer Treat Commun 2015;4:165–8. https://doi.org/10.1016/j.ctrc.2015.09.007.

[76] Zhou R, Xu T, Nguyen Q, Liu Y, Yang J, Komaki R. Radiation dose, local disease progression, and overall survival in patients with inoperable non-small cell lung cancer after concurrent chemoradiation therapy. Radiat Oncol Biol 2018;100:452–61. https://doi.org/10.1016/j.ijrobp.2017.10.003.

[77] Warner A, Dahele M, Hu B, Palma DA, Senan S, Oberije C, et al. Factors associated with early mortality in patients treated with concurrent chemoradiation therapy for locally advanced non-small cell lung cancer. Int J Radiat Oncol Biol Phys 2016;94:612–20. https://doi.org/10.1016/j.ijrobp.2015.11.030.

[78] Mao Q, Xia W, Dong G, Chen S, Wang A, Jin G, et al. A nomogram to predict the survival of stage IIIA-N2 non-small cell lung cancer after surgery 1784–1792.e3. J Thorac Cardiovasc Surg 2018;155. https://doi.org/10.1016/j.jtcvs.2017.11.098.

[79] Hwang IG, Ahn MJ, Park BB, Ahn YC, Han J, Lee S, et al. ERCC1 expression as a prognostic marker in N2(+) nonsmall-cell lung cancer patients treated with platinum-based neoadjuvant concurrent chemoradiotherapy. Cancer 2008;113:1379–86. https://doi.org/10.1002/cncr.23693.

[80] Lee VHF, Chan WWL, Lee EYP, Choy TS, Ho PPY, Leung DKC, et al. Prognostic significance of standardized uptake value of lymph nodes on survival for stage III non-small cell lung cancer treated with definitive

concurrent chemoradiotherapy. Am J Clin Oncol Cancer Clin Trials 2016;39:355–62. https://doi.org/10.1097/COC.0000000000000070.

[81] Dieleman EMT, Uitterhoeve ALJ, Van Hoek MW, Van Os RM, Wiersma J, Koolen MGJ, et al. Concurrent daily cisplatin and high-dose radiation therapy in patients with stage III non-small cell lung cancer. Int J Radiat Oncol Biol Phys 2018;102:543–51. https://doi.org/10.1016/j.ijrobp.2018.07.188.

[82] Yoo GS, Oh D, Pyo H, Ahn YC, Noh JM, Park HC, et al. Concurrent chemoradiotherapy for unresectable non-small cell lung cancer invading adjacent great vessels on radiologic findings: is it safe ? J Radiat Oncol 2019;60:234–41. https://doi.org/10.1093/jrr/rry102.

[83] Pusceddu C, Melis L, Sotgia B, Guerzoni D, Porcu A, Fancellu A. Usefulness of percutaneous microwave ablation for large non-small cell lung cancer: A preliminary report. Oncol Lett 2019;18:659–66. https://doi.org/10.3892/ol.2019.10375.

[84] Konert T, Vogel WV, Paez D, Polo A, Fidarova E, Carvalho H, et al. Introducing FDG PET/CT-guided chemoradiotherapy for stage III NSCLC in low- and middle-income countries: preliminary results from the IAEA PERTAIN trial. Eur J Nucl Med Mol Imaging 2019;46:2235–43. https://doi.org/10.1007/s00259-019-04421-5.

[85] Kim D, Song C, Kim SH, Kim YJ, Lee JS, Kim J. Chemoradiotherapy versus radiotherapy alone following induction chemotherapy for elderly patients with stage III lung cancer. Radiat Oncol J 2019;37:176–84.

[86] Chaudhuri MR. Primary pulmonary cavitating carcinomas. Thorax 1973;28:354–66. https://doi.org/10.1136/thx.28.3.354.

[87] National Comprehensive Cancer Network. NCCN clinical practive guidelines in oncology non-small cell lung cancer. Version 3. 2018.

[88] Mountain CF. Staging of lung cancer. Yale J Biol Med 1981;54:161–72.

[89] Goldstraw P, Chansky K, Crowley J, Rami-porta R, Asamura H, Eberhardt WEE, et al. The IASLC lung cancer staging poject: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. J Thorac Oncol 2015;11:39–51. https://doi.org/10.1016/j.jtho.2015.09.009.

[90] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJ, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 2020;295:328–38. https://doi.org/10.1148/radiol.2020191145.

[91] Welch ML, McIntosh C, Haibe-kains B, Milosevic MF, Wee L, Dekker A, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol 2019;130:2–9. https://doi.org/10.1016/j.radonc.2018.10.027.

[92] Van Zandwijk N. Neoadjuvant strategies for non-small cell lung cancer. Lung Cancer 2001;34:S145–50. https://doi.org/10.1016/S0169-5002(01)00359-2.

[93] Meko J, Rusch VW. Neoadjuvant therapy and surgical resection for locally advanced non-small cell lung cancer. Semin Radiat Oncol 2000;10:324–32.

CHAPTER 3

The association between muscle quantity and overall survival depends on muscle radiodensity: a cohort study in non-small cell lung cancer patients

Wouter A.C. van Amsterdam[1,2,3,*], Netanja I. Harlianto[1],
Joost J.C. Verhoeff[3] , Pim Moeskops[4], Pim A. de Jong[1], Tim Leiner[1,5]

[1] University Medical Center Utrecht, department of Radiology,
[2] Babylon Health,
[3] University Medical Center Utrecht, department of Radiation Oncology,
[4] Quantib BV,
[5] Mayo Clinic, department of Radiology

Journal of Personalized Medicine (2022) 12(7), 1191
https://doi.org/10.3390/jpm12071191

The prognostic value of CT-derived muscle quantity for overall survival (OS) in patients with Non-Small Cell Lung cancer (NSCLC) is uncertain due to conflicting evidence. We hypothesize that increased muscle quantity is associated with better OS in patients with normal muscle radiodensity but not in patients with fatty degeneration of muscle tissue and low muscle radiodensity.

We performed an observational cohort study in NSCLC patients treated with radiotherapy. A deep learning algorithm was used to measure muscle quantity as psoas muscle index (PMI) and psoas muscle radiodensity (PMD) on computed tomography. The potential interaction between PMI and PMD for OS was investigated using Cox proportional-hazards regression. Baseline adjustment variables were age, sex, histology, performance score and body mass index. We investigated non-linear effects of continuous variables and imputed missing values using multiple imputation.

We included 2840 patients and observed 1975 deaths in 5903 patient years. The average age was 68.9 years (standard deviation, 10.4, range 32 to 96) and 1692 (59.6%) were male. PMI was more positively associated with OS for higher values of PMD (hazard ratio for interaction 0.915; 95% confidence interval 0.861 – 0.972; p-value 0.004).

We found evidence that high muscle quantity is associated with better OS when muscle radiodensity is higher, in a large cohort of NSCLC patients treated with radiotherapy. Future studies on the association between muscle status and OS should accommodate this interaction in their analysis for more accurate and more generalizable results.

Keywords:

Cachexia; Carcinoma, Non-Small-Cell Lung; survival analysis; prognosis

Abbreviations

Overall survival (OS)
Non-Small Cell Lung Cancer (NSCLC)
Computed tomography (CT)
Psoas muscle index (PMI)
Skeletal muscle index (SMI)
Psoas muscle radiodensity (PMD)
Skeletal muscle radiodensity (SMD)
Electronic health records (EHR)
Positron Emission Tomography (PET)
American Joint Committee on Cancer Tumor-Node-Metastasis (TNM) staging protocol
Hounsfield Units (HU)
Intravenous Contrast (IV)
Body mass index (BMI)
Standard Deviation (SD)

Introduction

Skeletal muscle quantity is related to patient prognosis in Non-Small Cell Lung Cancer (NSCLC) for several reasons. First, muscle loss occurs more frequently in patients with aggressive tumors with high catabolic activity [1]. Second, patients who have low muscle mass are less capable of enduring intensive anti-cancer treatments such as surgery [2], chemotherapy [3] and radiotherapy [4].

Standardized measurements of muscle quantity on computed tomography (CT) scans are the psoas muscle index (PMI) and the skeletal muscle index (SMI). PMI is defined as the cross-sectional area of the psoas muscle on the CT slice corresponding to lumbar vertebra 3 (L3), divided by the square of the height of a patient [5]. SMI is analogously defined by taking the cross-sectional area of all skeletal muscle on the L3 slice instead of only the psoas muscle. Studies in NSCLC patients report that higher muscle quantity correlates with improved Overall Survival (OS) [6–8], improved progression-free survival [9] and better response to treatment [10]. However, most of these results represent univariable associations and several studies did not find that muscle quantity was correlated with these outcomes [1, 11–13]. In addition to muscle quantity, the radiodensity of muscle tissue has gained interest as a potential prognostic marker in NSCLC as well [11–17]. Pro-inflammatory cytokines are elevated in cancer patients and lead to fatty infiltration of muscle tissue. Fatty infiltration can be measured on a CT scan as fat tissue has a lower radiodensity than muscle tissue. Standardized measurements of fatty muscle infiltration are the average radiodensity in the psoas muscle area on the L3 level, known as psoas muscle radiodensity (PMD), and the analogous average radiodensity in all skeletal muscle on L3, skeletal muscle radiodensity (SMD).

PMD and SMD provide measurements of the extent of fatty infiltration of muscle tissue. When muscle tissue is replaced by intra-muscular adipose tissue, muscle strength decreases. We therefore hypothesize that in patients with high muscle radiodensity, muscle quantity is more positively associated with OS than in patients with low muscle radiodensity. This would imply that there is a statistical interaction between muscle quantity and muscle radiodensity for OS. Whereas there is a large body of research on the prognostic value of muscle quantity and muscle radiodensity for OS in lung cancer separately (see for example these systematic reviews: [18, 19]), few studies have investigated the associations of both muscle quantity and muscle radiodensity with OS [11–17]. Whether the association between muscle quantity and OS depends on muscle radiodensity has not been studied before. If the association between muscle quantity and OS indeed depends on muscle radiodensity, the apparent association between muscle quantity and OS will differ across studies if their study populations differ in distribution of muscle radiodensity. If our hypothesis is true, OS prediction models that include the interaction between muscle quantity and muscle radiodensity will be more accurate and more stable across populations. Therefore, we investigated whether there is a statistical interaction between muscle quantity and muscle radiodensity for OS prediction in a large cohort of NSCLC patients treated with radiotherapy.

Materials and Methods

Data source

We conducted a retrospective observational cohort study at the department of radiotherapy of the University Medical Center Utrecht. Patients were referred to our center from 9 different hospitals in the Utrecht province, the Netherlands. This study was conducted in accordance with the applicable privacy guidelines and the declaration of Helsinki and its later amendments. As this was a retrospective study and most of the patients had died, we obtained a waiver for informed consent from the institutional review board at the University Medical Center Utrecht (reference number WAG/mb/19/005583).

Patient inclusion

We identified and included patients if they had visited the radiotherapy department for consideration of treatment with radiotherapy for NSCLC between January 2009 and September 2018. Some patients had multiple episodes of NSCLC for which they received radiotherapy. For these patients we only included the first episode. Apart from this, there were no exclusion criteria, meaning that all patients were included in the final analysis.

Definition of exposures and outcome

Clinical variables were extracted from the electronic health records (EHR). The outcome for this study was OS measured on a continuous time scale. The start of follow-up was the date of the first visit to the radiotherapy department. This is generally when treatment decisions are made and where prognostic models have the highest potential impact. If a date of death was not registered in the health records, the Dutch Personal Records Database was queried to verify survival status. The last date of follow-up was April 26th, 2021.

Patients were staged according to the American Joint Committee on Cancer Tumor-Node-Metastasis (TNM) staging protocol. We maintained the TNM version that was clinically used at the time of treatment, spanning versions six [20], seven [21] and eight [22]. Positron Emission Tomography – Computed Tomography (PET-CT) scans made within the time window of 90 days before follow-up start to 30 days after follow-up start were used for the body composition measurements. We used PET-CT scans for our measurements as thoracic CT scans generally do not contain the required L3 level. If there were multiple eligible PET-CT scans for a patient, the scan that was closest to the start of follow-up was used. The PET-CT scans were made over a 9-year period and in 9 different hospitals, which means there was a natural variation in scanner vendor, model, tube current, voxel spacing, slice thickness and radiation dose. We used Quantib Body Composition version 0.2.1 (Quantib BV, Rotterdam, the Netherlands) for automated muscle measurements [23]. The CT scans were first resampled to a uniform slice thickness of 5mm. The first step for the algorithm was to select the CT slice in the middle of the third lumbar vertebra. On this slice, as well as the two slices above and the two slices below this center slice, the psoas muscle tissue was automatically segmented bilaterally. The cross-sectional area of the segmentation was measured in centimeters squared, and subsequently averaged over the five segmented slices covering a range of 2.5 cm. To calculate the psoas muscle area, only voxels with a radiodensity of -30 Hounsfield Units (HU) or higher were counted to exclude intra-muscular fatty tissue. The psoas muscle area was divided by the square of the length of a patient measured in meters to obtain the PMI. As a second measurement, the PMD was measured as the mean HU value in the entire segmented region, including voxels with radiodensity lower than -30 HU [12, 16]. The definitions of PMI and PMD are illustrated in Figure 1. PET-CT scans are rarely acquired with intravenous (IV) iodinated contrast, but if a scan was obtained after IV contrast injection only the PMI measurement was used, as iodinated contrast artificially increases radiodensity of muscle tissue. All automated segmentations were verified for correctness in joint reading sessions by three experienced readers (WA, NH, TL) including a board-certified radiologist with over 15 years of clinical experience (TL). If the automated segmentation failed, it was corrected manually by one of the authors (NH). The process of creation, verification and correction of segmentations was blinded to the outcome of OS and other patient information.

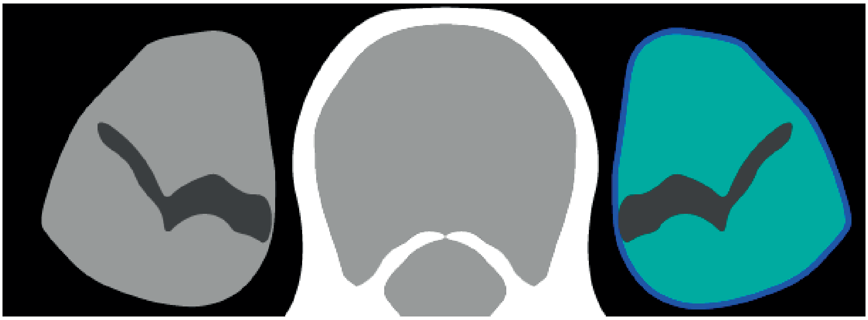


Figure 1. Schematic representation of measurements. The entire area of the psoas muscle on the L3 level was delineated (dark blue circumference). For psoas muscle index (PMI), only voxels with a radiodensity of -30 hounsfield units HU or higher were counted (light green area). For psoas muscle radiodensity (PMD), the average HU of all voxels in the delineated area was calculated, including fatty infiltration of the psoas muscle.

The following baseline clinical characteristics were extracted from the EHR: age, sex, histology group (grouped as adenocarcinoma, squamous cell carcinoma, no histology obtained or other), performance score defined by the Eastern Cooperative Oncology Group [24] and body mass index (BMI), defined as patient weight in kilograms divided by the square of the length in meters. These variables were selected based on their wide availability in clinical practice and their frequent inclusion in prognostic models. As patients may lose weight because of their NSCLC, the time window for weight measurements was 90 days before to 30 days after the start of follow-up.

Statistical analysis

Model definition

We used Cox proportional-hazards regression to model OS. As the purpose of this study was to evaluate the potential added prognostic value of the interaction between PMI and PMD, the baseline clinical covariates were included in all tested models. Continuous variables were centered by subtracting the mean before entering the analysis. Potential non-linear effects of the continuous predictors (age, BMI, PMI and PMD) were investigated by including restricted cubic spline terms using 5 knots which leads to 4 degrees of freedom per variable. The potential multiplicative interaction on the hazard ratio scale between PMI and PMD was investigated by including interaction terms. Only interaction terms that were linear in either PMI or PMD were included to reduce the number degrees of freedom needed to model the interaction. The average OS between different clinical disease stages is very different and the proportional hazards assumption is unlikely to hold for clinical stage. Therefore, the baseline hazard function was stratified per clinical stage in four groups (I, II, III and IV). The patient selection mechanism for radiotherapy is different for early-stage (I and II) and advanced-stage (III and IV) NSCLC. For early-stage NSCLC, patients without contra-indications for surgery are recommended for surgical treatment [25, 26]. For stage III, radiotherapy is a standard part of treatment [25, 26]. For stage IV, potential indications for radiotherapy are aggressive local treatment of oligometastastatic disease [25] or palliative care on a case-by-case basis [25]. As in early-stage NSCLC the treatment selection is dependent on their fitness for surgery, the treatment choice is likely correlated with their muscle quantity and radiodensity. As the patient selection mechanism differs between early-stage and advanced-stage, we stratified the hazard ratios per early-stage (stages I and II) and advanced-stage (stages III and IV). The full model included 58 parameters in total.

Sample size calculation

We used simulations to calculate the power to detect a hazard ratio of 0.986 for a linear interaction term between PMI and PMD for several different sample sizes and correlations between covariates. The assumptions for the sample size calculations were based on three published studies [12, 14, 17]. The simulations indicated that 1000 patients were sufficient for a power of 0.8 using a two-sided Student's T-test with alpha = 0.05 for a wide range of correlations between variables. The appendix presents a detailed report on the assumptions and results of the sample size calculation.

Missing data

The presence or absence of a PET-CT scan for a patient in our study depends on medical decisions made during the diagnostic and treatment planning process. It is likely that these decisions are correlated with the clinical variables under study and the outcome OS. This means that excluding all patients with missing data ('complete-case analysis') would lead to biased parameter estimates [27]. Given the baseline clinical variables included in our study and the outcome OS, the assumption of missing at random conditional on these variables may be tenable. In this situation, multiple imputation yields unbiased parameter estimates and increases the statistical power as more patients are included in the analysis [28]. Therefore, missing data in both the baseline clinical covariates and the scan-derived muscle measurements PMI and PMD were imputed using multiple imputation. To accommodate the non-linear dependencies between covariates and survival implied by the Cox proportional-hazards model, the non-linear terms of the continuous predictors and the interaction terms, we performed imputation using Substantive Model Compatible Fully Conditional Specification [29]. This ensures compatibility between the imputation models and the outcome model. Data were imputed under the most comprehensive outcome model under study.

Hypothesis testing

We compared models using the multi-parameter pooled Wald-test that is compatible with multiple imputation [30]. We tested two variants of our main hypothesis that there is a statistical interaction between PMI and PMD: 1. including non-linear interaction terms between PMI and PMD, and stratification per early-stage versus advanced-stage (14 degrees of freedom); 2. only a linear interaction between PMI and PMD without stratification (1 degree of freedom).

Implementation

R version 4.1.0 was used for all statistical analyses. The function 'smcfcs' from package smcfcs (version 1.5.0) was used for imputation. Data were imputed using 250 iterations per imputation and 160 fully imputed datasets were generated. The function D1 from package MICE (version 3.13.0) was used for multi-parameter model comparisons. The function rcspline.eval from package Hmisc (version 4.5.0) was used to generate the restricted cubic spline bases for continuous variables. To accommodate the mixed stratification of baseline hazards and hazard ratios we updated the source code of packages smcfcs and survival. The code that implements the imputation and subsequent analysis is publicly available here: https://doi.org/10.5281/zenodo.6107815.

Reporting

For reporting, we adhered to the REMARK statement for biomarker studies [31]. A filled form is available in the supplemental material.

Results

We included 2840 patients and observed 1975 deaths in 5903 patient years. The average age was 68.9 years (standard deviation, 10.4, range 32 to 96) and 1692 (59.6%) were male. The median OS since first visit to the radiotherapy department ranged from 3.32 years for stage I patients to 0.53 years for stage IV patients. The baseline characteristics stratified per clinical stage are presented in Table 1 and per-stage Kaplan-Meier survival curves are presented in the appendix (Appendix Figure 1).

Table 1. Baseline characteristics stratified by clinical disease stage.

	Overall	**stage I**	**stage II**	**stage III**	**stage IV**	**missing**
n	**2840**	**714**	**145**	**871**	**343**	**767**
age (mean (SD))	68.95 (10.44)	72.65 (9.18)	71.63 (10.47)	66.53 (10.24)	66.26 (10.21)	68.97 (10.73)
male sex (%)	1692 (59.6)	422 (59.1)	89 (61.4)	531 (61.0)	211 (61.5)	439 (57.2)
histology (%)						
adenocarcinoma	595 (21.0)	81 (11.3)	32 (22.1)	272 (31.2)	136 (39.7)	74 (9.6)
no examination	1402 (49.4)	482 (67.5)	55 (37.9)	190 (21.8)	83 (24.2)	592 (77.2)
other	259 (9.1)	46 (6.4)	13 (9.0)	121 (13.9)	56 (16.3)	23 (3.0)
squamous cell	508 (17.9)	74 (10.4)	43 (29.7)	278 (31.9)	59 (17.2)	54 (7.0)
missing	76 (2.7)	31 (4.3)	2 (1.4)	10 (1.1)	9 (2.6)	24 (3.1)
PS (%)						
0	872 (30.7)	177 (24.8)	23 (15.9)	206 (23.7)	64 (18.7)	402 (52.4)
1	553 (19.5)	154 (21.6)	29 (20.0)	239 (27.4)	61 (17.8)	70 (9.1)
>=2	446 (15.7)	102 (14.3)	31 (21.4)	153 (17.6)	80 (23.3)	80 (10.4)
missing	969 (34.1)	281 (39.4)	62 (42.8)	273 (31.3)	138 (40.2)	215 (28.0)
BMI (mean (SD))	25.66 (6.07)	25.57 (5.96)	25.57 (5.26)	25.73 (5.64)	26.42 (7.75)	25.32 (6.23)
BMI missing (%)	1500 (52.8)	309 (43.3)	64 (44.1)	417 (47.9)	212 (61.8)	498 (64.9)
PMI (mean (SD))	6.28 (1.64)	6.27 (1.74)	6.09 (1.41)	6.41 (1.59)	6.21 (1.74)	43.59 (8.43)
PMI missing (%)	1851 (65.2)	386 (54.1)	79 (54.5)	525 (60.3)	266 (77.6)	595 (77.6)
PMD (mean (SD))	27.93 (10.89)	25.81 (12.28)	26.99 (11.32)	30.99 (9.21)	29.33 (10.07)	7.16 (13.88)
PMD missing (%)	1637 (57.6)	314 (44.0)	68 (46.9)	442 (50.7)	262 (76.4)	551 (71.8)
RT target (%)						
lung	1520 (53.5)	667 (93.4)	92 (63.4)	179 (20.6)	126 (36.7)	456 (59.5)
multi-site	1040 (36.6)	29 (4.1)	37 (25.5)	618 (71.0)	146 (42.6)	210 (27.4)
other	114 (4.0)	12 (1.7)	7 (4.8)	16 (1.8)	31 (9.0)	48 (6.3)
mediastinum	97 (3.4)	5 (0.7)	0 (0.0)	43 (4.9)	19 (5.5)	30 (3.9)
hilus	37 (1.3)	0 (0.0)	7 (4.8)	11 (1.3)	5 (1.5)	14 (1.8)
thoraxwall	23 (0.8)	1 (0.1)	2 (1.4)	4 (0.5)	8 (2.3)	8 (1.0)
brain	8 (0.3)	0 (0.0)	0 (0.0)	0 (0.0)	7 (2.0)	1 (0.1)
missing	1 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.3)	0 (0.0)
SBRT (%)	1096 (38.6)	643 (90.1)	61 (42.1)	29 (3.3)	39 (11.4)	324 (42.2)
deceased (%)	1975 (69.5)	364 (51.0)	96 (66.2)	674 (77.4)	284 (82.8)	557 (72.6)
survival (median)	1.71	3.32	2.15	1.41	0.53	1.63

The mean and standard deviation are calculated based on the non-missing values. The 'other' category for histology includes carcinoid tumors, neuro-endocrine tumors and other rare histologic subtypes. Variables age, male sex, SBRT, deceased and survival time had no missing values. A comprehensive dedicated table of radiotherapy targets is presented in the appendix (Figure 4). Median overall survival was calculated using the Kaplan-Meier method. PS: performance score, defined using the Eastern Collaborative Oncology Group standard [24]. SD: standard deviation. BMI: body mass index, PMI: psoas muscle mass index, PMD: psoas muscle radiodensity, SBRT: stereotactic body radiation therapy, RT: radiotherapy.

For 1212 of the 2840 patients (42.7%) a PET-CT scan was available within the required time window and muscle measurements were performed. Only one of these PET-CT scans was with intravenous contrast. The median number of days from the scan to the start of follow-up was 33 (interquartile range 21-49). We used 10 variables per patient in the analysis (age, sex, histology group, performance score, BMI, PMI, PMD, clinical stage, survival time, deceased indicator) meaning that there were 28,400 potential values to be recorded. Of these 28,400 values, 21,600 were observed and 6,800 were missing meaning that 76% of the data were available and 24% were imputed. There were 378 patients (13.3%) with no missing values for any of the variables. If a complete-cases analysis were employed, 3,780 out of 21,600 (17.5%) available data-points would be used, disregarding 82.5% of the available data.

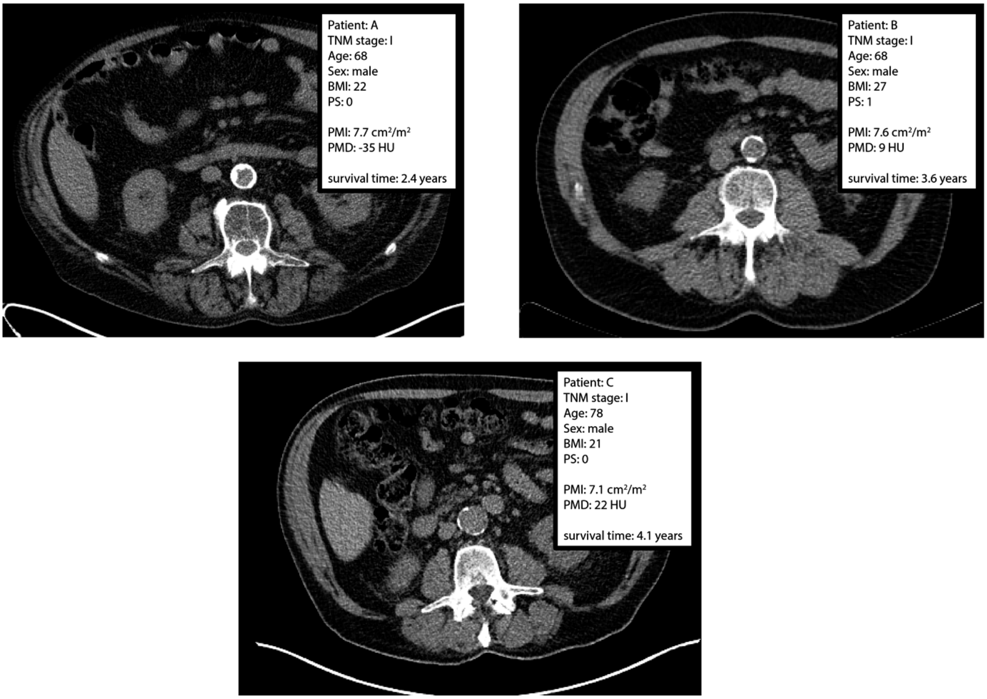


Figure 2. L3-slices of computed tomography scans for three different patients with similar psoas muscle index (PMI) but different psoas muscle radiodensity (PMD). BMI: body mass index, PS: performance score, defined using the Eastern Collaborative Oncology Group standard [24].

Three patients with similar PMI but different PMD are presented in Figure 2. The dependence of the association between PMI and OS on PMD is presented in Figure 3. The shape of the interaction curve confirms the hypothesis that PMI is more positively associated with OS when PMD is higher. In other words, increased psoas muscle area (PMI) is associated with increased OS only when psoas muscle radiodensity (PMD) is sufficiently high. Accordingly, there was clear statistical evidence for a linear interaction (hazard ratio 0.915; 95% confidence interval 0.861 – 0.972; p-value 0.004). There was no statistical evidence for non-linear components and stratification of the interaction (p-value 0.667). A table with the parameter estimates for all parameters is presented in the appendix (Table 5).

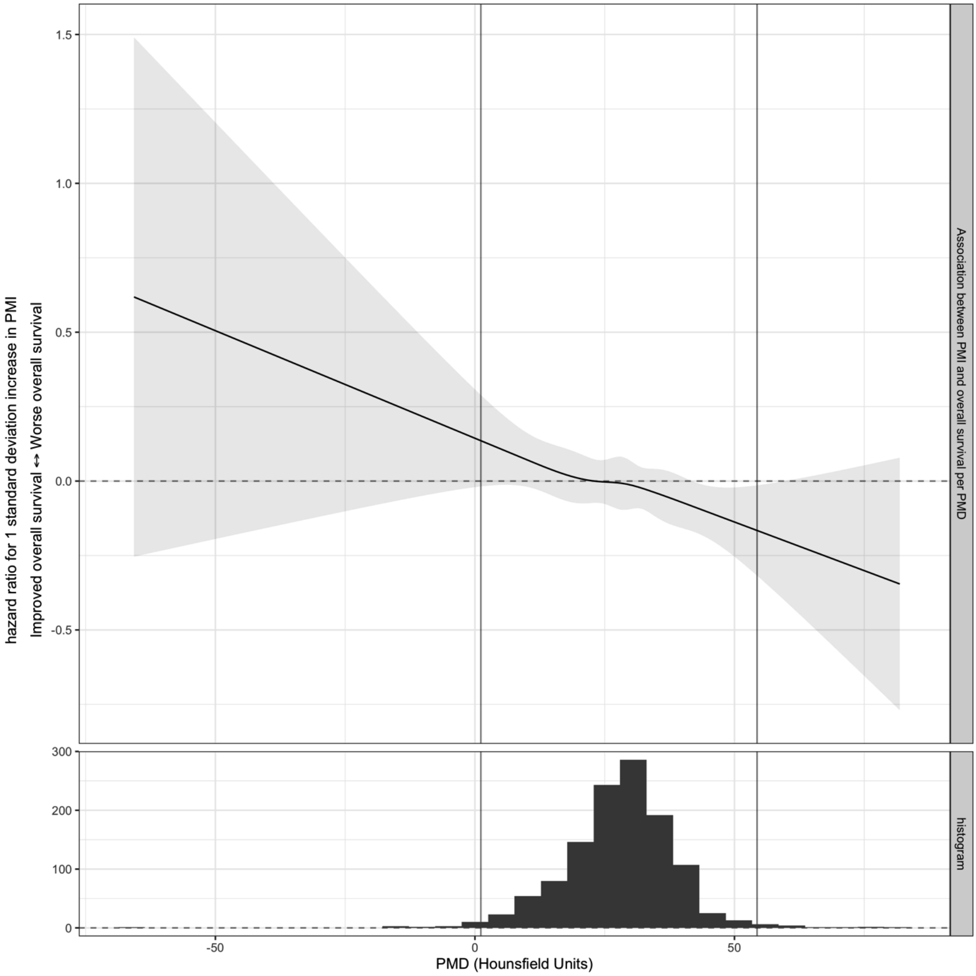


Figure 3. Hazard ratio for a 1 standard deviation increase in PMI for different values of PMD. The average estimate is depicted with a solid black line. The 95% confidence interval is depicted with the gray shaded area. The dashed line indicates the null effect of hazard ratio 1. At the bottom a histogram for the observed values of PMD is presented. Two vertical lines indicate the region excluding the 1% lowest and 1% highest values of PMD. For this figure, the model was fitted by omitting non-linear terms of PMI and stratification of hazard ratios per early-stage versus advanced-stage. The full model also includes non-linear interaction terms of PMI which means that the shape of this interaction function also depends on the value of PMI. To estimate the confidence interval, the model was fitted on 100 bootstrap samples of each of the 160 imputed datasets, following the 'MI-boot' procedure [32]. PMI: skeletal muscle index, PMD: skeletal muscle radiodensity.

Discussion

We conducted a large cohort study in non-small cell lung cancer patients treated with radiotherapy and investigated whether the relationship between muscle quantity and overall survival (OS) depends on muscle radiodensity. Our experiments confirmed the hypothesis that the higher the muscle radiodensity measured in PMD, the more positive the association between muscle quantity measured in PMI and OS. These findings provide a potential explanation for the varying results in previous research on muscle mass and OS and have important implications for future research. Future studies on muscle mass and OS should accommodate this statistical interaction between muscle quantity and muscle radiodensity.

The study cohort consisted of a heterogeneous group of NSCLC cancer patients treated with radiotherapy over a span of 9 years. Stage I and II NSCLC patients treated with radiotherapy are known to have worse overall survival than stage I and II patients treated with surgery because most patients treated with radiotherapy were deemed unfit for surgery. Indeed, the overall survival of stage I and II patients in our population is lower than the general population [33] but similar to other radiotherapy-only populations [34, 35]. Given the biological rationale for the hypothesis that muscle quantity of sufficient radiodensity is more protective for OS than muscle quantity that is infiltrated with fatty tissue, we suspect that this hypothesis is true across all cancer types, stages and treatment regimes. The clear statistical evidence in a heterogeneous population of NSCLC cancer patients supports this suspicion, but it will have to be confirmed in future studies in multiple cancer types.

Our study has several limitations. Although our cohort is relatively large, we do not have extensive details on the included patients, specifically with respect to other potential treatments they received. If our aim were to present a new prediction model for use in clinical practice, this lack of detail would be an important limitation. Instead, our goal was to evaluate a hypothesis that has implications for future prediction research. As it is unlikely that the interaction between muscle quantity and muscle radiodensity depends on the given treatment, our findings are meaningful despite the lack of detail on other treatments. Finally, there were relatively few patients with complete data. In accordance with statistical guidelines [36, 37] we used state-of-the art imputation methods to optimally use the information that was available without excluding any of the patients. In total 24% of the data points were imputed. Multiple imputation remains valid even when there are many missing values as long as a sufficient number of imputed datasets is used [38]. Still, future studies should preferable be based on prospective cohorts where important variables are collected in a protocolled manner. Another way to investigate our hypothesis further is by conducting a systematic review of studies on the association between muscle area and muscle radiodensity with OS in cancer patients and performing a meta-regression of the association between muscle area and OS on muscle radiodensity.

In conclusion, we found that PMI is more positively associated with overall survival when PMD is higher in a large cohort of NSCLC patients treated with radiotherapy. For accurate and generalizable results, future studies on the relationship between muscle quantity and overall survival in cancer patients should accommodate this statistical interaction in the analysis.

Acknowledgements

Disclosure statement

All authors declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work; PJ received consulting fees from Sanfit and Inozyme. TL is co-founder and clinical advisor of Quantib BV, PM is employee of Quantib BV. The department of radiology at the University Medical Center Utrecht has a research collaboration with Quantib BV and Philips Healthcare.

Funding

WA was supported by the Alexandre Suerman personal PhD stipendium. The Alexandre Suerman stipend had no role in any part of the study design, conduct or reporting. After this research was conducted but before final publication, WA started working at Babylon Health Limited, United Kingdom.

References

[1] Srdic D, Plestina S, Sverko-Peternac A, et al (2016) Cancer cachexia, sarcopenia and biochemical markers in patients with advanced non-small cell lung cancer—chemotherapy toxicity and prognostic value. Supportive Care in Cancer 24:4495–4502

[2] Baracos V, Kazemi-Bajestani SMR (2013) Clinical outcomes related to muscle mass in humans with cancer and catabolic illnesses. The International Journal of Biochemistry & Cell Biology 45:2302–2308. https://doi.org/10/gmgzx2

[3] Ryan AM, Prado CM, Sullivan ES, et al (2019) Effects of weight loss and sarcopenia on response to chemotherapy, quality of life, and survival. Nutrition 67–68:110539. https://doi.org/10/gmgzxq

[4] Topkan E, Parlak C, Topuk S, Pehlivan B (2012) Influence of oral glutamine supplementation on survival outcomes of patients treated with concurrent chemoradiotherapy for locally advanced non-small cell lung cancer. BMC Cancer 12:502. https://doi.org/10.1186/1471-2407-12-502

[5] Mourtzakis M, Prado CMM, Lieffers JR, et al (2008) A practical and precise approach to quantification of body composition in cancer patients using computed tomography images acquired during routine care. Appl Physiol Nutr Metab 33:997–1006. https://doi.org/10/bshwnj

[6] Portal D, Hofstetter L, Eshed I, et al (2019) L3 skeletal muscle index (L3SMI) is a surrogate marker of sarcopenia and frailty in non-small cell lung cancer patients. Cancer Management and Research 11:2579–2588

[7] Shoji F, Matsubara T, Kozuma Y, et al (2017) Relationship between preoperative sarcopenia status and immuno-nutritional parameters in patients with early-stage non-small cell lung cancer. Anticancer Research 37:6997–7003

[8] Suzuki Y, Okamoto T, Fujishita T, et al (2016) Clinical implications of sarcopenia in patients undergoing complete resection for early non-small cell lung cancer. Lung Cancer 101:92–97

[9] Takada K, Yoneshima Y, Tanaka K, et al (2020) Clinical impact of skeletal muscle area in patients with non-small cell lung cancer treated with anti-PD-1 inhibitors. Journal of Cancer Research and Clinical Oncology 146:1217–1225

[10] Stene GB, Helbostad JL, Amundsen T, et al (2015) Changes in skeletal muscle mass during palliative chemotherapy in patients with advanced lung cancer. Acta Oncologica 54:340–348

[11] Nattenmüller J, Wochner R, Muley T, et al (2017) Prognostic impact of CT-quantified muscle and fat distribution before and after first-line-chemotherapy in lung cancer patients. PLoS ONE 12:. https://doi.org/10.1371/journal.pone.0169136

[12] Sjøblom B, Grønberg BH, Wentzel-Larsen T, et al (2016) Skeletal muscle radiodensity is prognostic for survival in patients with advanced non-small cell lung cancer. Clinical Nutrition 35:1386–1393. https://doi.org/10.1016/j.clnu.2016.03.010

[13] Nishioka N, Naito T, Notsu A, et al (2021) Unfavorable impact of decreased muscle quality on the efficacy of immunotherapy for advanced non-small cell lung cancer. Cancer Med 10:247–256. https://doi.org/10.1002/cam4.3631

[14] Abbass T, Dolan RD, MacLeod N, et al (2020) Comparison of the prognostic value of MUST, ECOG-PS, mGPS and CT derived body composition analysis in patients with advanced lung cancer. Clinical Nutrition ESPEN 40:349–356. https://doi.org/10/gjv5r8

[15] Bowden JCS, Williams LJ, Simms A, et al (2017) Prediction of 90 Day and Overall Survival after Chemoradiotherapy for Lung Cancer: Role of Performance Status and Body Composition. Clinical Oncology 29:576–584. https://doi.org/10.1016/j.clon.2017.06.005

[16] Cortellini A, Bozzetti F, Palumbo P, et al (2020) Weighing the role of skeletal muscle mass and muscle density in cancer patients receiving PD-1/PD-L1 checkpoint inhibitors: a multicenter real-life study. Scientific Reports 10:

[17] Dolan RD, Maclay JD, Abbass T, et al (2020) The relationship between 18F-FDG-PETCT-derived tumour metabolic activity, nutritional risk, body composition, systemic inflammation and survival in patients with lung cancer. Scientific Reports 10:. https://doi.org/10.1038/s41598-020-77269-7

[18] Shachar SS, Williams GR, Muss HB, Nishijima TF (2016) Prognostic value of sarcopenia in adults with solid tumours: A meta-analysis and systematic review. European Journal of Cancer 57:58–67. https://doi.org/10.1016/j.ejca.2015.12.030

[19] Takenaka Y, Oya R, Takemoto N, Inohara H (2021) Predictive impact of sarcopenia in solid cancers treated with immune checkpoint inhibitors: a meta-analysis. J Cachexia Sarcopenia Muscle 12:1122–1135. https://doi.org/10.1002/jcsm.12755

[20] TNM Classification of Malignant Tumours, 6th Edition. In: Wiley.com. https://www.wiley.com/en-nl/TNM+Atlas%2C+6th+Edition-p-9781118695609. Accessed 1 Dec 2020

[21] TNM Classification of Malignant Tumours, 7th Edition. In: Wiley.com. https://www.wiley.com/en-nl/TNM+Classification+of+Malignant+Tumours%2C+7th+Edition-p-9781444358964. Accessed 1 Dec 2020

[22] TNM Classification of Malignant Tumours, 8th Edition. In: Wiley.com. https://www.wiley.com/en-us/TNM+Classification+of+Malignant+Tumours%2C+8th+Edition-p-9781119263579. Accessed 4 Dec 2020

[23] van Erck D, Moeskops P, Schoufour J, et al (2022) Evaluation of a fully automatic deep learning-based method for the measurement of psoas muscle area. Frontiers in Nutrition

[24] Oken MM, Creech RH, Tormey DC, et al (1982) Toxicity and response criteria of the Eastern Cooperative Oncology Group. American Journal of Clinical Oncology 5:649–656

[25] Ettinger DS (2020) NCCN Non-Small Cell Lung Cancer Guideline, Version 1.2021. https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf. Accessed 15 Feb 2021

[26] NVALT Niet kleincellig longcarcinoom - Resectabel en lokaal uitgebreid NSCLC - Richtlijn - Richtlijnendatabase

[27] Little RJA, Rubin DB (2002) Complete-Case and Available-Case Analysis, Including Weighting Methods. In: Statistical Analysis with Missing Data. John Wiley & Sons, Ltd, pp 41–58

[28] Little RJA, Rubin DB (2002) Estimation of Imputation Uncertainty. In: Statistical Analysis with Missing Data. John Wiley & Sons, Ltd, pp 75–93

[29] Bartlett JW, Seaman SR, White IR, Carpenter JR (2015) Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Stat Methods Med Res 24:462–487. https://doi.org/10.1177/0962280214521348

[30] Li KH, Raghunathan TE, Rubin DB (1991) Large-Sample Significance Levels from Multiply Imputed Data Using Moment-Based Statistics and an F Reference Distribution. Journal of the American Statistical Association 86:1065–1073. https://doi.org/10.2307/2290525

[31] McShane LM, Altman DG, Sauerbrei W, et al (2005) REporting recommendations for tumour MARKer prognostic studies (REMARK). Br J Cancer 93:387–391. https://doi.org/10.1038/sj.bjc.6602678

[32] Schomaker M, Heumann C (2018) Bootstrap Inference When Using Multiple Imputation. Stat Med 37:2252–2266. https://doi.org/10.1002/sim.7654

[33] Kay FU, Kandathil A, Batra K, et al (2017) Revisions to the Tumor, Node, Metastasis staging of lung cancer (8th edition): Rationale, radiologic findings and clinical implications. World Journal of Radiology 9:269–279. https://doi.org/10.4329/wjr.v9.i6.269

[34] Yerokun BA, Yang C-FJ, Gulack BC, et al (2017) A national analysis of wedge resection versus stereotactic body radiation therapy for stage IA non–small cell lung cancer. The Journal of Thoracic and Cardiovascular Surgery 154:675-686.e4. https://doi.org/10.1016/j.jtcvs.2017.02.065

[35] Rowell NP, Williams C (2001) Radical radiotherapy for stage I/II non-small cell lung cancer in patients not sufficiently fit for or declining surgery (medically inoperable). Cochrane Database of Systematic Reviews. https://doi.org/10.1002/14651858.CD002935

[36] Moons KGM, Altman DG, Reitsma JB, et al (2015) Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med 162:W1. https://doi.org/10/gfrkkz

[37] Altman DG, McShane LM, Sauerbrei W, Taube SE (2012) Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. BMC Med 10:51. https://doi.org/10/gb33js

[38] von Hippel PT (2020) How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. Sociological Methods & Research 49:699–718. https://doi.org/10.1177/0049124117747303

Appendix

Sample size simulations

We used simulations to calculate the required sample size. Performing these simulations requires assumptions on three things: the outcome model (i.e. hazard ratios and baseline hazard function), the correlation between predictor variables and the censoring distribution.

As Sjøblom et al. (1) present the most complete multivariable analysis, we use their results (presented in their Table 3) as the basis for assumed parameter values. The assumed parameter values are presented in Table 1. The assumed parameter value for our target parameter SMISMD (the linear interaction between skeletal muscle index, SMI, and skeletal muscle radiodensity, SMD) was taken to be halfway between the hazard ratio for SMI and the hazard ratio for SMD on the log-hazard ratio scale. As the interpretation of the absolute value of a hazard ratio relies on the scale of the variable, the hazard ratio for the interaction term was rescaled by multiplying with the standard deviation of SMD and dividing by the standard deviation of SMISMD in each simulated dataset. The hazard ratios for histology subtypes were not given so we assumed values for these.

Table 1. Hazard ratios for sample size calculations.

Term	hazard_ratio	log_hazard_ratio
Age	0.99	-0.010
Male sex	0.77	-0.261
Histology: other	1.22	0.2
Histology: squamous	1.35	0.3
BMI	0.99	-0.01
PS 1	1.24	0.215
PS >=2	1.89	0.636
SMD	0.98	-0.017
SMI	0.99	-0.010
SMISMD	0.99	-0.014

Adenocarcinoma is the reference category for histology group. BMI: body mass index, PS: ECOG performance score (0 is the reference category), SMD: skeletal muscle radiodensity, SMI: skeletal muscle index, SMISMD: interaction term between SMI and SMD.

For all variables, marginal statistics (mean and standard deviations for continuous variables, frequency tables for discrete variables) were extracted from (1). As a frequency table for the four NSCLC stages was not available from (1), we used two additional publications to reconstruct the frequency table for clinical stage. Dolan et al. provided the relative frequencies of stages I, II and III (2). Abbass et al. provided relative frequencies for stages III and IV (3). These relative frequencies were used to reconstruct a single full frequency table for all four stages. In addition to the hazard ratios for the individual parameters, the power also depends on the correlation between the predictor variables. As a complete covariance matrix for all variables was not available, we simulated covariate data using covariance structures induced by different Clayton copulas (4). Copulas are multivariate cumulative distribution functions whose marginal distributions are uniform on the unit interval. A Clayton copula can be defined using the known marginal statistics of the observed variables and a single unknown correlation parameter. This can be done by translating the marginal distributions (assumed to be Gaussian for continuous variables, binomial for binary variables and discretized Gaussian for discrete variables) of the variables to cumulative distribution functions. The inverse of these cumulative distribution functions are also uniform on the unit interval by definition and can then be identified with the marginal distributions from the copula. The relationship between the Clayton copula parameter and the average Pearson correlation coefficient of variables generated from such a copula is presented in Table 2.

Table 2. Clayton copula parameter versus average Pearson correlation coefficient of two variables simulated by a Clayton copula with that parameter value.

Copula parameter	Pearson correlation
0.1	0.043
0.2	0.115
0.3	0.231
0.4	0.258
0.5	0.326
0.6	0.398
0.7	0.386
0.8	0.408
0.9	0.486
1	0.510

Finally, the power also depends on the marginal survival distributions and the censoring distributions. We estimated the marginal survival distributions per stage, and the censoring distribution for all stages from our data. We used the parametric power generalized Weibull model (5) to estimate these survival distributions and to simulate survival times. To prevent extreme outliers with high leverage, simulated survival times over 15 years were censored. Kaplan-Meier estimates and power generalized Weibull estimates of the marginal survival distributions per stage are presented in Figure 1.

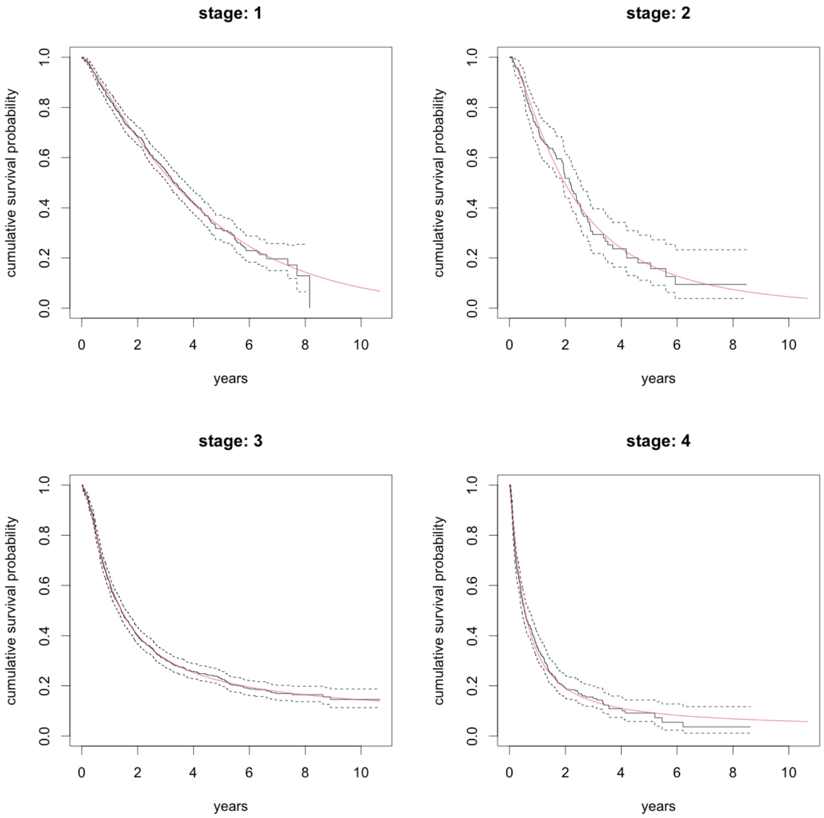


Figure 1. Marginal survival distributions per stage. The Kaplan-Meier estimate is presented with the solid black line, accompanied by a 95% confidence interval indicated with the dotted black line. The parametric power generalized Weibull estimate that was used in the simulations is indicated with the red line.

We calculated the power to detect the pre-specified interaction hazard ratio at a 0.05 significance level using a two-sided Student's T-test. We calculated the power over the following grid of values: Copula parameter 0.1, 0.25, 1.0; sample size 500, 1000, 2000. For each of the 9 combinations we simulated 1000 datasets. The power was defined as the number of times a significant result was detected divided by the total number of simulations for that setting. The results of the power analysis are presented in Table 3.

Table 3. Results of power analysis for the interaction term between skeletal muscle index (SMI) and skeletal muscle radiodensity (SMD).

Power	Sample size	Copula parameter
0.576	500	0.1
0.551	500	0.25
0.517	500	1
0.887	1000	0.1
0.865	1000	0.25
0.791	1000	1
0.99	2000	0.1
0.992	2000	0.25
0.968	2000	1

Supplemental tables

Table 4. Overview of different target regions for radiotherapy per stage.

target	missing	stage I	stage II	stage III	stage IV
missing	0	0	0	0	1
brain	1	0	0	0	7
hilus	14	0	7	11	5
hilus, supraclavicular	1	0	0	0	1
lung	456	667	92	179	126
lung, hilus	19	3	7	9	3
lung, mediastinum	158	26	28	571	121
lung, mediastinum, hilus	3	0	0	16	1
lung, mediastinum, supraclavicular	5	0	0	8	2
lung, supraclavicular	3	0	0	3	2
lung, thoraxwall	4	0	2	2	7
lung, thoraxwall, vertebra	1	0	0	0	0
lung, vertebra	2	0	0	1	0
mediastinum	30	5	0	43	19
mediastinum, hilus	13	0	0	5	4
mediastinum, hilus, thoraxwall	0	0	0	1	0
mediastinum, hilus, vertebra	0	0	0	0	1
mediastinum, supraclavicular	0	0	0	1	2
mediastinum, vertebra	0	0	0	0	1
other	43	12	7	15	28
plexus	0	0	0	0	1
supraclavicular	2	0	0	0	1
thoraxwall	8	1	2	4	8
thoraxwall, vertebra	1	0	0	1	1
vertebra	3	0	0	1	1

Table of parameter estimates

Table 5. Estimates of all parameters in the full model with linear interaction term and without stratification of the interaction.

term	estimate	std.error	statistic	df	p.value
age	-0.018	0.122	-0.144	1329.330	0.886
sex_maleTRUE	0.188	0.097	1.929	713.392	0.054
histono_pa	0.002	0.127	0.015	316.727	0.988
histoother	0.194	0.106	1.831	1525.555	0.067
histosquamous	0.208	0.089	2.336	1524.444	0.020
ecog_bin1	0.099	0.115	0.857	338.536	0.392
ecog_bin2	0.511	0.113	4.515	381.416	0.000
bmi	0.043	0.216	0.198	325.775	0.843
smi	-0.243	0.184	-1.315	345.824	0.189
smd	0.053	0.197	0.268	501.070	0.789
age1	0.183	0.535	0.342	1339.720	0.733
age2	0.828	3.376	0.245	1349.396	0.806
bmi1	-1.802	1.685	-1.069	439.175	0.286
bmi2	12.337	8.705	1.417	485.333	0.157
smi1	1.127	1.323	0.852	405.681	0.395
smi2	-2.792	6.181	-0.452	413.667	0.652
smd1	-0.712	0.713	-0.998	509.379	0.319
smd2	5.390	5.772	0.934	537.145	0.351
age3	-3.845	7.203	-0.534	1345.819	0.594
bmi3	-18.958	11.953	-1.586	514.767	0.113
smi3	0.486	8.643	0.056	409.716	0.955
smd3	-7.733	11.220	-0.689	554.587	0.491
smismd	-0.089	0.031	-2.868	391.247	0.004
age:c_stage_earlyTRUE	-0.160	0.288	-0.556	463.172	0.579
sex_maleTRUE:c_stage_earlyTRUE	0.031	0.142	0.218	709.628	0.828
histono_pa:c_stage_earlyTRUE	0.006	0.200	0.033	538.576	0.974
histoother:c_stage_earlyTRUE	-0.008	0.242	-0.032	1228.706	0.974
histosquamous:c_stage_earlyTRUE	0.184	0.196	0.939	1240.738	0.348
ecog_bin1:c_stage_earlyTRUE	-0.506	0.190	-2.659	281.168	0.008
ecog_bin2:c_stage_earlyTRUE	-0.417	0.185	-2.257	326.862	0.025
bmi:c_stage_earlyTRUE	-0.089	0.337	-0.264	244.152	0.792
smi:c_stage_earlyTRUE	-0.017	0.292	-0.059	267.759	0.953
smd:c_stage_earlyTRUE	-0.171	0.222	-0.770	611.886	0.441
age1:c_stage_earlyTRUE	0.498	1.004	0.496	762.491	0.620
age2:c_stage_earlyTRUE	-3.642	5.699	-0.639	902.736	0.523
bmi1:c_stage_earlyTRUE	-0.798	2.690	-0.297	335.899	0.767
bmi2:c_stage_earlyTRUE	3.673	14.236	0.258	361.337	0.797
smi1:c_stage_earlyTRUE	1.101	2.166	0.508	299.204	0.612
smi2:c_stage_earlyTRUE	-6.375	10.000	-0.638	320.489	0.524
smd1:c_stage_earlyTRUE	0.162	0.949	0.171	480.186	0.865
smd2:c_stage_earlyTRUE	2.547	8.788	0.290	422.612	0.772
age3:c_stage_earlyTRUE	7.489	11.143	0.672	1000.686	0.502
bmi3:c_stage_earlyTRUE	-3.758	19.964	-0.188	372.093	0.851
smi3:c_stage_earlyTRUE	9.752	13.834	0.705	333.289	0.481
smd3:c_stage_earlyTRUE	-9.027	18.269	-0.494	406.953	0.621

The estimates are provided on the log-hazard ratio scale. All continuous variables are scaled to unit variance. Higher order cubic spline terms of continuous variables are not scaled to unit variance which explains the otherwise extremely high parameter estimates. ECOG performance score 0 is the reference category for ecog_bin1 and ecog_bin2. Histology type adenocarcinoma is the reference category for the other histology types.

References

[1] Sjøblom B, Grønberg BH, Wentzel-Larsen T, Baracos VE, Hjermstad MJ, Aass N, et al. Skeletal muscle radiodensity is prognostic for survival in patients with advanced non-small cell lung cancer. Clinical Nutrition. 2016;35(6):1386–93.

[2] Dolan RD, Maclay JD, Abbass T, Colville D, Buali F, MacLeod N, et al. The relationship between 18F-FDG-PETCT-derived tumour metabolic activity, nutritional risk, body composition, systemic inflammation and survival in patients with lung cancer. Scientific Reports [Internet]. 2020;10(1). Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096909045&doi=10.1038%2fs41598-020-77269-7&partnerID=40&md5=dedc04a0cc3d872c6f11ab44a6e4bb0d

[3] Abbass T, Dolan RD, MacLeod N, Horgan PG, Laird BJ, McMillan DC. Comparison of the prognostic value of MUST, ECOG-PS, mGPS and CT derived body composition analysis in patients with advanced lung cancer. Clinical Nutrition ESPEN. 2020;40:349–56.

[4] Yan J. Enjoy the Joy of Copulas: With a Package copula. Journal of Statistical Software. 2007 Oct 8;21(1):1–21.

[5] Burke K, Jones MC, Noufaily A. A Flexible Parametric Modelling Framework for Survival Analysis. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2020 Apr;69(2):429–57.

PART II

Improving treatment decisions with individual treatment effect estimates

CHAPTER 4

Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning

W. A. C. van Amsterdam[1], J. J. C. Verhoeff[2], P. A. de Jong[1], T. Leiner[1] and M. J. C. Eijkemans[3]

[1] Department of Radiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.
[2] Department of Radiation Oncology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.
[3] Biostatistics and Research Support, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.

npj Digital Medicine (2019) 2:122
https://doi.org/10.1038/s41746-019-0194-x

Abstract

Deep learning has shown remarkable results for image analysis and is expected to aid individual treatment decisions in health care. Treatment recommendations are predictions with an inherently causal interpretation. To use deep learning for these applications in the setting of observational data, deep learning methods must be made compatible with the required causal assumptions. We present a scenario with real-world medical images (CT-scans of lung cancer) and simulated outcome data. Through the data simulation scheme, the images contain two distinct factors of variation that are associated with survival, but represent a collider (tumor size) and a prognostic factor (tumor heterogeneity), respectively. When a deep network would use all the information available in the image to predict survival, it would condition on the collider and thereby introduce bias in the estimation of the treatment effect. We show that when this collider can be quantified, unbiased individual prognosis predictions are attainable with deep learning. This is achieved by (1) setting a dual task for the network to predict both the outcome and the collider and (2) enforcing a form of linear independence of the activation distributions of the last layer. Our method provides an example of combining deep learning and structural causal models to achieve unbiased individual prognosis predictions. Extensions of machine learning methods for applications to causal questions are required to attain the long-standing goal of personalized medicine supported by artificial intelligence.

Introduction

Deep learning has many possible applications in health care, especially for tasks including unstructured data such as medical images. Convolutional neural networks (CNN) are deep learning models that have demonstrated remarkable performance on many tasks including images. These models are attractive for prediction tasks on medical images, as CNNs can be optimized end-to-end from image to outcome. This way the network can detect patterns in the images that are relevant to the prediction task, but may be unknown to medical professionals. A downside is that the induced representations of the network are 'hidden' and not readily interpretable. A much sought after holy grail of artificial intelligence is to attain personalized treatment decisions through individual prognosis prediction and individual treatment effect estimation. Treatment effect estimation is a causal question, so answering it requires techniques from causal inference.[1] A pivotal result from causal inference is that when the direction of causal relationships between variables in a given situation is known, identifiability and estimands of causal queries can be deduced automatically using do-calculus. In the case of treatment effect estimation of lung cancer measured with overall survival, this means that we must know (a) which variables affect both treatment allocation and overall survival, (b) the causal direction of relationships between the variables. For instance, we know that the level of pre-treatment overall fitness is related to the likelihood of getting intensive treatment. In this case the direction of causation is clear due to the time ordering: pre-treatment fitness influences the treatment decision, and not vice versa. Whether this is a strong or weak relationship, or the specific functional form of the relationship (e.g., whether the relationship is monotonic) is not important for the consideration of general non-parametric causal effect identification. These causal relationships can be encoded succinctly in a Directed Acyclic Graph (DAG) with an arrow pointing from the cause to the effect, e.g., fitness → treatment. When the DAG that encodes the relationship between all the relevant variables is known, do-calculus provides an answer to whether a specific causal question can be answered from the observed data.

The connection between images and a DAG is not always straightforward to see. Fundamentally, patient outcomes are driven by biological processes, and images may contain (more or less noisy) views of these processes. For example, a particularly aggressive lung tumor may grow very large, as can be seen on CTscans, and this biological behavior leads to worse overall survival. These biological processes can be seen as underlying causes of factors of variation or patterns in the image in the language of structural causal models. Conversely, information derived from medical images is often used to make treatment decisions. Here, the image is a causal factor for treatment selection. When a deep neural network is used to predict a certain clinical outcome, it will make use of all factors of variation in an image that are statistically associated with that outcome. Thus, predicting an outcome with deep learning based on an image can be seen as conditioning on (noisy views of) the underlying causal factors of the patterns in these images. Medical images, especially images from large body parts such as a chest CT-scan in the case of lung cancer, may contain many different factors of variation that can have different 'roles' in the DAG. Notably when a specific factor of variation represents a collider in the DAG, conditioning on the image by using a deep learning model may introduce bias in the estimation of treatment effects

A collider is a variable that is the effect of two or more variables. To explain collider bias, consider the following clinical scenario. The pulmonary oncology department in a general hospital serves the population of a small geographic region for all cases of lung cancer, and 90% of their patients come from this region. However, one of the oncologist has a special interest in the treatment of a rare form of lung cancer: carcinoid tumors, accounting for roughly 1% of lung cancer cases. Everyone in the country with this rare form of lung cancer visits this single specialist for their treatment. Being treated in this hospital for lung cancer is a collider, as it has two causes: living in the surrounding region, or having the rare carcinoid form. In reality, these two causes are independent: the risk of getting carcinoid lung cancer is the same for everyone, regardless of the region of residence. However, within the population of the patients treated in this hospital there appears to be a strong inverse relationship between living in this specific region and having carcinoid lung cancer. Patients who are treated in the hospital but are not from the surrounding region are very likely to have the rare form, whereas patients who live close to the hospital are very unlikely to have carcinoid lung cancer (namely 1%). This observed 'spurious' correlation is the result of conditioning on a collider through restricting the patient sample to only those treated in this single hospital. Including an indicator for being treated in this hospital as a regression variable in a multiinstitutional study into lung cancer is another form of conditioning that will lead to similar collider bias.

We describe a fictional but realistic clinical scenario where the following conditions hold: (1) There exists a clinical need for outcome prediction. (2) This outcome partly depends on treatment, and an unbiased estimate of the treatment effect is required. (3) The DAG describing the data-generating process is assumed to be known. (4) An image is hypothesized to contain important information for the task in (1), however, one of the factors of variation in the image represents a collider in the DAG. Conditioning on this collider will lead to a biased estimate of (2). (5) The collider can be measured from the image. (6) Deep learning is used to optimally predict (1). We stress that this poses a conflicting problem: 'simply' using deep learning to predict the outcome based on the image may lead to a low prediction error of the outcome in the observed data, but it will lead to bias in the estimated effect of treatment, as it conditions on a collider. No matter how accurate the resulting predictions are on the observed data, such models cannot accurately predict in the setting where we intervene on treatment. This effectively nullifies the clinical usefulness of the model for selecting the best treatment for new patients. The model only 'works' when treatments are allocated as was always done without the model. On the other hand, ignoring the image all together will lead to worse prediction error as the image contains important prognostic information. Our contribution is that we show that by utilizing a multi-task prediction scheme for both the outcome and the collider, accompanied by an additional loss term to induce a form of linear independence between final layer activations, we can satisfy both (1) the supervised prediction task and (2) attain an unbiased estimate of the treatment effect. For clarity in notation, we will reserve the term prediction error for performance on the supervised prediction task (e.g., accuracy of predicted survival time). With bias we will refer to difference between the expectation of the estimated treatment effect and the data-generating mechanism.

Results

Clinical case

The proposed clinical case concerns the treatment of lung cancer. Optimal treatment selection for lung cancer patients is a challenging problem: depending on the clinical disease stage, patients receive (combinations of) chemotherapy, radiotherapy, surgery, or more recently, immunotherapy or targeted therapy.[2] Some patients will be cured, while others only endure invalidating side-effects. In addition to using disease stage, personalized treatment decisions may be aided by estimating the individual prognosis of a patient for the different modes of treatment that are available. Medical scans provide important information for diagnosing and staging lung cancer, but may also provide this prognostic information. Deep learning is particularly attractive to analyze these scans, as these models may discover new prognostic factors or treatment effect modifiers.

Data-generating mechanism

In our experiments we use a public data set of chest CT-scans from the Lung Image Database Consortium image collection (LIDC[3]) These 1018 scans from 1010 unique patients each contain lung nodules ($N = 2609$) suspected of lung cancer. Up to four radiologists segmented the nodules on each consecutive image slice. As described in the original publication of the data, the data where gathered from seven participating hospitals and the study was approved by the appropriate local institutional review boards (IRB). Informed consent procedures were followed according to local IRB guidelines, and the data collection and anonymization were conducted in compliance with the Health Insurance Portability and Accountability Act (HIPAA) guidelines with the intent of providing a publicly available data set. Our study is conducted in accordance with the usage guidelines from the data provider.[4] We do not add new patient data, so IRB approval for this specific study was not needed. A CT-scan measures radiodensity, and tissues may exhibit different density-patterns. Heterogeneity in radiodensity is known to be associated with higher biologic aggressiveness and worse survival.[5] We used nodule size and the variance of radiodensity in a simulation study involving a binary treatment and a real-valued outcome reflecting overall survival. Note that our simulation does not accurately reflect the real world. Real world applications would require more complex models. The aim of our contribution is to address a current limitation in methodological tools. Therefore we chose the simplest graphical model that induces the problem we try to solve, but is still clinically conceivable. A DAG used for a real-world clinical application will be much more complex, but may still include the basic collider structure we present in this simulation and will thus require similar methods. Figure 1 and Table 1 illustrate the following hypothetical narrative.

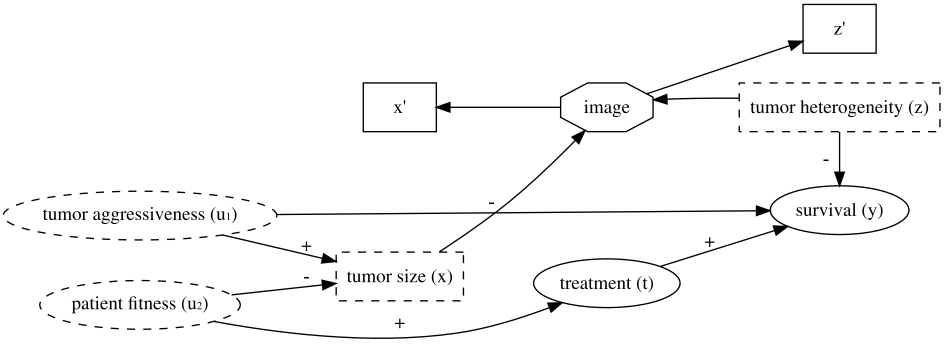


Figure 1. Directed Acyclic Graph describing the data-generating mechanism for the simulations. Signs indicate positive or negative associations. Rectangle shaped variables are image variables, dashed variables are unobserved. Tumor aggressiveness and patient fitness cannot be directly measured. x, z represent biological processes, causing the outcome

and image patterns. We cannot directly observe these biological processes, but x' , z' are noisy views of these variables that are measurable from the image. x is a collider since it is the child of u_1 and u_2. Conditioning on x will induce an artificial association between u_1 and u_2, thereby inducing a confounding path between treatment and survival, that only exists when conditioning on the collider.

There exist two possible treatments for lung cancer: $t \in \{0,1\}$, where $t = 1$ is deemed more aggressive and also more effective. An unobserved variable u_2 influences treatment allocation: people who appear to be in better overall health, as per subjective judgment of the physician, will have a higher probability of being treated with $t = 1$. At the same time these fitter patients generally have a better functioning immune system. The immune system combats the lung cancer, leading to a lower tumor size (x). Another unobserved variable u_1 represents the tumor biologic aggressiveness. High aggressiveness leads to a bigger tumor and negatively impacts the overall survival. We emphasize that the tumor size (x) is a pre-treatment collider according to this causal graph. A third noise variable, heterogeneity of radiodensity (z), is a prognostic factor unrelated to the treatment, but related to the outcome. Tumors with high heterogeneity lead to reduced survival.

Table 1. Parameters for sampling images and modeling outcome data.

	Variable	Variable model
u_1	Aggressiveness	$N(0, 0.7071)$
u_2	Fitness	$N(0, 0.7071)$
z	Heterogeneity	$N(0, 1)$
x	Size	$N(u_1 - u_2; 0.05)$
t	Treatment	$Bern(invlogit(N(u_2 - 0.5, 0.05)))$
y	Survival	$N(t - z - 2u_1 - 0.5; 0.05)$

For each observation i, an image is drawn from the total pool of images with the closest x_i and z_i. This ensures the required association between factors of variation in the image and the simulated outcome data. The parametric equations follow the DAG presented in Fig. 1: u_1 , u_2 , z are continuous independent noise variables. The collider x is the difference between u_1 and u_2, with a small amount of Gaussian noise (standard deviation of noise = 0:05). u_1 and u_2 have a standard deviation of $0.7071 \approx \sqrt{2}/2$ to ensure that x has a standard deviation of ≈ 1. Treatment t is modeled as a Bernoulli variable with a logistic link function, where increased u_2 increases the probability of being treated. 0.5 is subtracted to assure that ~50% of patients are treated. Gaussian noise of standard deviation 0.25 is added to the inverse log-odds of being treated to assure that every patient has some probability of being treated with the more intense treatment. This reflects the clinical world better as some patients may have strong preferences regarding their treatment, regardless of their underlying health status. Overall survival (y) increases with treatment (the true treatment effect is 1) and decreases with heterogeneity in radiodensity and tumor aggressiveness. Again, Gaussian noise of standard deviation 0.05 is added to introduce some uncertainty in the data.

This situation leads to a conundrum. As can be seen from the DAG, the marginal average treatment effect is identified by ATE = $E[p(y|t = 1) - p(y|t = 0)]$. The conditional treatment effect is not identified when conditioning the entire image, which is a descendant of both x and z. Conditioning on x' (the tumor size as measured in the image), corresponds to partly conditioning the collider x. This will induce an artificial association between u_1 and u_2, thereby opening a confounding path from t to y and violating of the backdoor criterion.[1] A backdoor path is a path from treatment to outcome that starts in the non-causal direction (an arrow pointing to the treatment instead of away from). This is indicative of confounding. When all confounding variables can be measured and conditioned on, all backdoor paths can be 'closed' during analysis, and the treatment effect can still be identified from observational data. In this case, a new backdoor path is introduced by conditioning on x', a proxy of x. This new path runs through the unobserved variables $t \leftarrow u_1 - u_2 \rightarrow y$. Therefore it cannot be closed by conditioning on these variables in the estimation, and the treatment effect is no longer identified. Using a convolutional neural network to predict y without regard for the biasing effect of conditioning on the collider will lead to a biased estimate of the treatment effect. Disentangling the factors of variation in the image to only utilize image information that is not related to the collider would enable an unbiased estimate of the conditional treatment effect, which is the goal of this study. The simulated data are visualized in Supplementary Figs 1 and 2.

Modeling

Our method, as summarized in Fig. 2, revolves around two central notions: (1) Utilizing the resemblance of the final layer of a CNN with linear regression and (2) Separating the contributions of different factors of variation during training to enable exclusion of factors of variation after model convergence. For each patient we have two observed quantities: $y_i \in \mathbb{R}$ and $t_i \in \{0,1\}$, along with an image which contains noisy views (x'_i , $z'_i \in \mathbb{R}$) of the tumor size x_i and heterogeneity z_i . The tumor size x_i is known to be a collider and can be measured from the image, tumor heterogeneity z_i is an unknown prognostic factor that we expect a CNN can 'discover' by training it to predict survival. Following standard practice for predicting a continuous real outcome with deep learning, the last layer of the CNN resembles linear regression where $\hat{y} = \beta_0 + \beta_t t + \sum_{j=1}^{N_k} \beta_j^k a_j^k$, with a_j^k the N_k activations of the final layer of a *k*-layer CNN, *t* the binary treatment indicator and β_0 an overall intercept. Indices for patients are omitted for clarity. Note that β_t is the estimated average treatment effect (ATE). The standard minibatch mean squared error is used for *y*:

$$L_y = \frac{1}{m} \sum_{i=1}^{m} (\hat{y} - y)^2 \tag{1}$$

where m the minibatch size. To attain separation of the collider from other factors of variation in the last layer, we modify the loss function such that a single activation of the last layer will approximate the collider: ak $a_1^k \approx x$. At the same time we optimize the other last layer activations $\{a_j^k, j > 1\}$ to be linearly independent of *x*'. Note that this is a light constraint based on the prior knowledge represented in the DAG, namely that *x* is a scalar and *x* and *z* are independent. We argue that after model convergence, we can fix all CNN parameters and do a single ordinary least squares on $\{a_j^k \cup t, j > 1\}$ to get a valid estimate of the treatment effect with β_t . These activations are constrained to be linearly independent of the collider, so performing linear regression on these activations and the treatment indicator should mimic omitting the collider as a variable in the regression. To attain this, we add a loss term for the collider *x*':

$$L_x = \frac{1}{m} \sum_{i=1}^{m} (a_1^k - x')^2 \tag{2}$$

This encourages the model to have a single activation in the last layer that approximates the collider x. This loss is synergistic with L_y as predicting *x*' from the image will improve L_y since *x* and *y* are statistically associated. At each training step, a prediction $\hat{x}^{\text{reg}}$ is made by regressing *x*' on the remaining last layer activations $\{a_j^k, j > 1\}$ with ordinary least squares. The MSE of this regression measures how well *x*' can be predicted from a linear combination of the other last layer activations $\{a_j^k, j > 1\}$. This is compared with the MSE of predicting x'_i with $\bar{x}$, the mean of *x*' of that minibatch of patients. When predicting x'_i from $\{a_j^k, j > 1\}$ is no better than using the mean of *x*', these activations are sufficiently independent from *x*. When the converse is true, the difference in mean squared errors is added to the total loss.

$$L_{\text{reg}} := \max(0, \text{MSE}(\bar{x}, x') - \text{MSE}(\hat{x}^{\text{reg}}, x')) \tag{3}$$

The total loss is the direct sum of these losses.

$$L = L_y + L_x + L_{\text{reg}} \tag{4}$$

Training was continued until convergence or overfitting, as assessed by an increase in total loss on the independently simulated validation set with different images than in the training set. After convergence, all CNN parameters were fixed and the final layer activations were calculated for each image. A linear regression of y was fitted on $\{a_j, t; 1 < j \leq N_k\}$ using the training set, resulting in a final model dubbed 'CausalNet'.

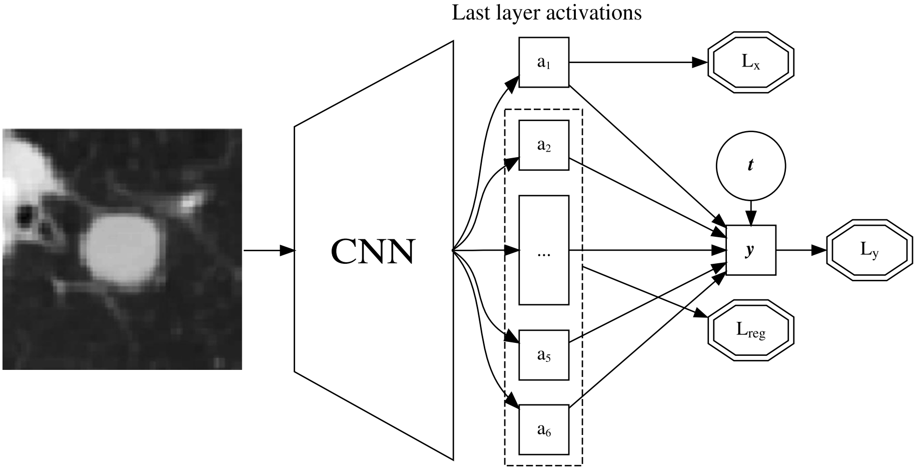


Figure 2. Schematic overview of the proposed convolutional neural network architecture. The network receives two inputs: an image and the treatment indicator (t). Loss functions are depicted in double octagons. The last layer activations are used to separate factors of variation in the image. a_1 is trained to approximate the measurement of the collider x'. The rest of the last layer activations are constrained to be linearly independent from x' through L_{reg}. The total loss is $L = L_y + L_x + L_{reg}$. CNN convolutional neural network.

Experiments

We calculated three baseline models for comparison: (1) ignoring all image information and using only the treatment indicator, (2) linear regression on the ground truth data $\{t, x, z, y\}$ with (2) and without (3) conditioning on the collider x. Through the sampling scheme, along with ambiguity in manual nodule segmentations and limitations of statistical learning from finite data, there is inherent prediction error for y and x. We estimated the MSE of this inherent error by predicting the ground truth labels x and z with a separate run of the same CNN architecture by replacing y with z. For fair comparison of the methods, in the regression baseline models we replaced x, z by x', z' by adding gaussian noise to the simulated x, z based on the MSE of the ground truth run for both variables. We compare the 'curve fitting' approach of conditioning on the entire image for predicting y (BiasNet) with the proposed method (CausalNet). As presented in Table 2, the proposed method separates the biasing effect of the collider x from the estimated treatment effect, and attains a prediction error close to the ideal expected loss for predicting y.

Table 2. Main results.

Model	**Variable**	**MSE$_y$**	**ATE**
Regression	t	2.74	1.02
Regression	t, x', z'	1.39	0.65
Regression*	t, z'	1.99	1.00
BiasedNet	t, *image*	1.83	0.66
CausalNet	$t, a_j^k (j > 1)$	2.23	1.02

Mean squared error for survival (MSE$_y$) along with estimated average treatment effect (ATE). The linear regression metrics are the expected outcomes according to whether or not the model conditions on the collider x. Regression* is the optimal value for our setup: (1) predicting the outcome based on relevant prognostic information from the image while (2) retaining a valid estimate of the treatment effect. All metrics were calculated on the validation set.

Measurement error

To test the sensitivity of our method to measurement error in the measured collider x, we simulated two additional scenarios where the collider was measured on the wrong scale. In one scenario, the actual relationship between the collider and the outcome was linear in the diameter of the nodule, while it was measured in units of volume. This represents a power 3 mismatch between the measurement and the actual relationship. The inverse scenario was studied as well. See Supplementary Fig. 3 for a visualization of relationship the measured x' and the true x. As shown in Table 3, the method seems robust to these kinds of measurement errors.

Table 3. Sensitivity analysis to measuring the collider on the wrong scale.

Model	Actual x	Measured x'	MSE_y	ATE
Regression*	Area	Area	1.99	1.00
CausalNet	Area	Area	2.23	1.02
CausalNet	Diameter	Volume	2.24	0.99
CausalNet	Volume	Diameter	2.21	1.02

Mean squared error for survival (MSE_y) along with estimated average treatment effect (ATE). The regression* results indicate the optimal results attainable for this simulated scenario.

Discussion

We provide a realistic medical example where plain curve fitting with deep learning will lead to biased predictions that do not generalize to the setting where we intervene on treatment. By utilizing prior knowledge about the world in the design of the CNN architecture and optimization scheme, accurate survival predictions were feasible with an unbiased estimate of the treatment effect. Our experiments demonstrate that deep learning can in principle be combined with insights from causal inference. Possible directions for extension of our experiments are introducing more elaborate data-generating mechanisms, for example with a treatment effect modifier or with statistical dependence between factors of variation within the image. In addition, similar approaches can be explored for medical images from different sources (e.g., pathology slides), or different data domains such as audio or natural language. We leave these extensions for further work.

Real world clinical applications of causal inference will necessarily involve more complicated DAGs. These DAGs could include one or more colliders. Our method can be adapted to multiple colliders in a straightforward manner by reserving a last layer activation for each collider, and requiring the other last layer activations to be independent of each of these colliders. Each realworld clinical scenario will require its own DAG for identifying treatment effects from observational data. Our contribution is that the proposed method can be used to attain deep representations of images that are independent of certain factors of variation.

Aside from the mitigation of collider bias, the proposed method can possibly be useful for other applications. For example, it may be used to produce deep representations of CT-scans that are independent of the scanner vendor. The scanner vendor would then take the place of the collider x in our simulation example.

To attain the goal of personalized treatment recommendations with artificial intelligence, methods combining machine learning with causal inference need to be further developed. Our experiments provide an example of how deep learning and structural causal models can be combined and are a small step forward towards personalized health care.

Methods

Data preparation and simulation

The LIDC-IDRI data set provides 1018 scans from 1010 patients with a total of 2609 nodules. The nodules were split in a training (70%) and validation (30%) set. Individual slices of the nodules were extracted and size (pixel count within segmentation) and heterogeneity (variance of pixel intensities within the segmented nodule) were calculated for each of the slices. Slices with a nodule size of <20 mm^2 were removed, as well as slices for which not all annotators agreed on the presence of a nodule. This yielded a training pool of 5015 slices and a validation pool of 1528 slices. Observations were simulated by sampling noise variables from the appropriate distributions and dependent variables according to the structural causal model in Table 1. For each patient i with simulated x_i , z_i , t_i , y_i , an image was drawn with replacement from the corresponding pool of images with the closest measured x' (size) and z' (heterogeneity). This sampling procedure induces a controllable statistical association between patterns in the image and the simulated treatment and outcome data. We simulated 3000 training observations and 1000 validation observations. Square slices of 7 × 7 cm surrounding the nodules were extracted from the CT-slices and resampled to isotropic 0.7 mm spacing. Pixel intensities were normalized to unit scale using a global mean and variance. The images were cropped randomly to 51 × 51 pixels during training, center crops of the same size were used for validation. In addition, random vertical and horizontal mirroring was used as data augmentation during training.

Neural network

We employed a VGG-like[6] CNN architecture. As our aim is to contrast methods of optimization for attaining unbiased predictions, we chose a simple CNN architecture with only basic layer types that was small enough for fast training but expressive enough to be able to model the nodule size and heterogeneity. The final network consisted 5 layers of 3 × 3 convolutions with 16 feature channels, each followed by a ReLU nonlinearity and 2 × 2 max-pooling. These basic image features were flattened into a 1 dimensional vector of size 144. Three fully connected layers of output sizes 144, 144, 12 were used, each followed by ReLU and dropout with $p = 0.25$, after which a final fully connected layer with output size $N_k = 6$ was used. The treatment indicator was concatenated to these activations for the final prediction during training. We used a batch size of 40 and the Adam optimizer[7] with a learning rate of 0.001 and no weightdecay.

References

[1] Pearl, J. *Causality* (Cambridge University Press, 2009).

[2] Ettinger, D. S. NCCN guidelines insights: non-small cell lung cancer, version 4.2016. *J. Natl. Compr. Canc. Netw.* **14**, 255–264 (2016).

[3] Armato, S. G. 3rd et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.* **38**, 915–931 (2011).

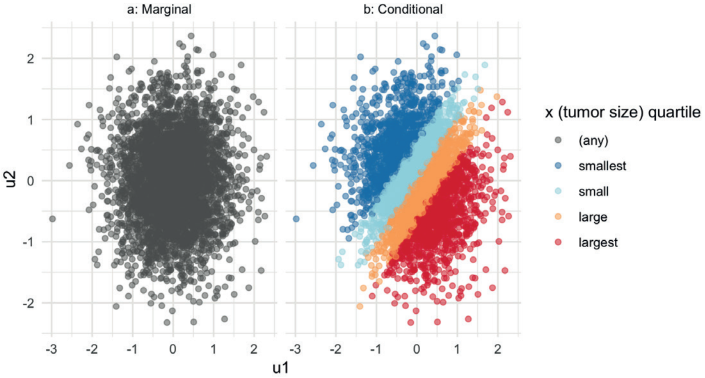
[4] Clark, K. et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).

[5] Bashir, U., Siddique, M. M., Mclean, E., Goh, V. & Cook, G. J. Imaging heterogeneity in lung cancer: techniques, applications, and challenges. *AJR Am. J. Roentgenol.* **207**, 534–543 (2016).

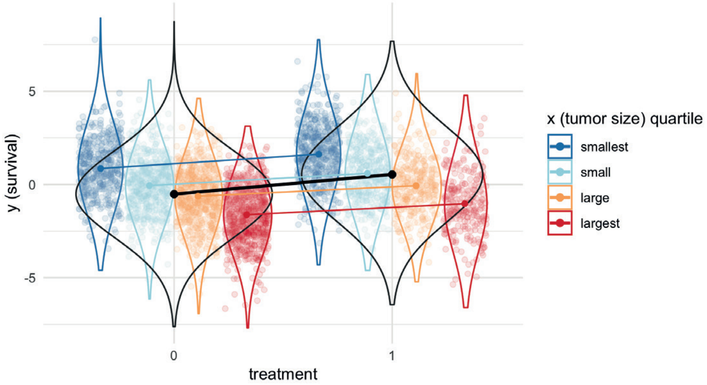
[6] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556 (2014).

[7] Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *arXiv preprint* arXiv:1412.6980 (2014).

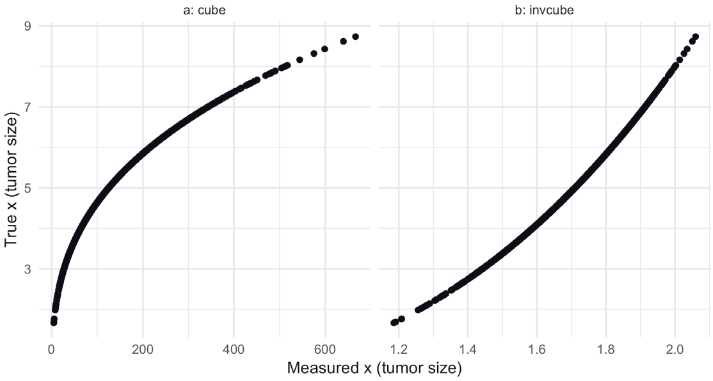
Supplemental Data:



Supplementary Figure 1: Visualization of the artificial correlation induced by conditioning on the collider x. Through the data generating mechanism, u_1 and u_2 are independently generated Gaussian variables. Facet a shows that the variables are marginally independent. The collider x is simulated as the difference between u_1 and u_2 with some Gaussian noise ($x \sim N(u_1 - u_2, 0.05)$). This means that for any given value of x, u_1 and u_2 are positively correlated. Facet b visualizes this artificial correlation by binning the values of x for the simulated data in quartiles.



Supplementary Figure 2: Visualization of the biasing effect of conditioning on the collider x (tumor size). The true treatment effect is 1. The solid black line visualizes the true (causal) difference in y (survival) between treated an untreated patients when **not** conditioning on the collider x. When observations are conditioned on the collider x, visualized here by binning patients in quartiles of x, the observed difference in survival between treated and untreated patients diminishes, as indicated by the colored lines. This diminished difference in survival between patients occurs when conditioning on the collider x. Conditioning on x induces a positive correlation between its parents u_1 and u_2. u_1 increases the probability of intensive treatment, while u_2 decreases the probability of survival. Due to this artificial association between u_1 and u_2, induced by conditioning on the collider x, the difference in survival between treated and untreated patients appears diminished. In reality, **assigning** a patient to intensive treatment will always increase their survival with 1, as reflected in the black line.



Supplementary Figure 3: Visualization of collider measurement and actual value. Facet a: x is measured as diameter, while y is linear in volume of the tumor Facet b: x is measured by volume, while y is linear in the diameter of the tumor.

CHAPTER 5

Individual treatment effect estimation in the presence of unobserved confounding using proxies: a cohort study in stage III non-small cell lung cancer

Wouter A.C. van Amsterdam[1,2*], Joost. J.C. Verhoeff[1], Netanja I. Harlianto[1], Gijs A. Bartholomeus[1], Aahlad Manas Puli[3], Pim A. de Jong[1], Tim Leiner[1], Anne S.R. van Lindert[1], Marinus J.C. Eijkemans[1], Rajesh Ranganath[3,4]

[1] University Medical Center Utrecht; Utrecht, the Netherlands
[2] Babylon Health; London, United Kingdom
[3] Courant Institute of Mathematical Sciences, New York University; New York, NY, United States
[4] Center for Data Science, New York University; New York, NY, United States
* corresponding author, w.a.c.vanamsterdam@gmail.com, work partly done at New York University

Scientific Reports 2022; 12(1):5848
https://doi.org/10.1038/s41598-022-09775-9

Abstract

Randomized Controlled Trials (RCT) are the gold standard for estimating treatment effects but some important situations in cancer care require treatment effect estimates from observational data. We developed “Proxy based individual treatment effect modeling in cancer” (PROTECT) to estimate treatment effects from observational data when there are unobserved confounders, but proxy measurements of these confounders exist. We identified an unobserved confounder in observational cancer research: overall fitness. Proxy measurements of overall fitness exist like performance score, but the fitness as observed by the treating physician is unavailable for research. PROTECT reconstructs the distribution of the unobserved confounder based on these proxy measurements to estimate the treatment effect. PROTECT was applied to an observational cohort of 504 stage III non-small cell lung cancer (NSCLC) patients, treated with concurrent chemoradiation or sequential chemoradiation. Whereas conventional confounding adjustment methods seemed to overestimate the treatment effect, PROTECT provided credible treatment effect estimates.

RCT: randomized controlled trial; PROTECT: Proxy based individual treatment effect modeling in cancer; NSCLC: non-small cell lung cancer

Introduction

Randomized controlled trials (RCT) are the gold standard for estimating treatment effects but there are important situations in cancer care where treatment effect estimates from observational data are needed. First, study participants of cancer RCTs are generally younger and in better overall health when compared to the real-world population *(1–3)*. Therefore, RCTs provide no direct evidence for applying the treatments in older and weaker patients, as these parts of the population are not covered by the trials. The effect of a treatment in these subpopulations could be estimated in observational data to investigate whether a RCT with extended inclusion criteria is indicated. Second, there is a constant research effort to discover new predictive biomarkers. Predictive biomarkers reveal parts of the biological behavior of the tumor that are related to the treatment effect and can be used to select the optimal treatment for a patient *(4)*. Before RCTs with new predictive biomarkers can be conducted, preliminary evidence on their association with the treatment effect is needed based on observational data.

Estimating treatment effects in observational data requires knowing what the confounders are of the treatment – outcome relationship. In cancer care, the most effective treatment is often also the most intensive treatment. The *overall fitness* of a patient determines what treatment intensity they can endure. Therefore, overall fitness is the central confounder. The treating physician will form an implicit assessment of overall fitness that is partly based on the subjective impression of a patient. As there is no record of this implicit assessment, traditional confounding adjustment methods cannot be used. However, proxy measurements of fitness are available, such as performance score. We developed a method named 'Proxy based individual treatment effect modeling in cancer' (PROTECT). PROTECT uses proxy measurements of fitness to reconstruct the distribution of the unobserved confounder and to adjust the treatment effect with this reconstructed confounder. In addition to modeling the confounder, PROTECT allows for incorporation of biomarkers of the biological behavior of the tumor. These biomarkers together with the patient overall fitness are used to predict the individual treatment effect.

We apply PROTECT to stage III Non-Small Cell Lung Cancer (NSCLC). Newly diagnosed patients with stage III NSCLC have two curative treatment options: concurrent chemotherapy and radiotherapy, or sequential treatment with chemotherapy followed by radiotherapy *(5)*. According to the meta-analysis of RCTs by Aupèrin et al., concurrent treatment leads to better overall survival (hazard ratio, 0.84; 95% confidence interval, 0.74 to 0.95) *(6)*. Concurrent treatment is more intensive and has a higher risk of severe toxicity *(6)*. Treatment guidelines recommend to give sequential treatment to patients with lower overall fitness *(5, 7)*. These patients are more likely to experience treatment toxicity under concurrent chemoradiation that would require adjustment or cessation of the treatment. This suggests that the survival benefit of concurrent treatment is absent or reversed in patients who are in lower overall fitness. As the real-world population contains older and weaker patients than the RCTs, an important question is whether the average treatment effect estimate from the RCTs is valid in the real-world population, and whether this average treatment effect applies to all patient subgroups.

In this study we present PROTECT as a method for estimating both the average treatment effect and individual treatment effects in real-world cancer populations from observational data. The method is applied to a multi-center observational cohort of stage III NSCLC patients, comparing concurrent chemoradiation with sequential chemoradiation.

Results

PROTECT

The objective of PROTECT is to estimate treatment effects from observational data. This requires knowing what the confounders are of the treatment – outcome relationship. In multiple discussion rounds with experts in oncology, thoracic oncology, radiotherapy, radiology, causal inference, statistics and epidemiology, we identified the confounders of the treatment – outcome relationship in cancer.

Patients who are in good overall health will more often be prescribed the most intensive and effective treatments as they can tolerate these treatments better than patients with lower overall fitness. Patients with better overall fitness also have better outcomes regardless of their treatment. This means that the overall fitness of a patient is a confounder of the treatment – outcome relationship. A treating physician will form an implicit assessment of the overall fitness of a patient that is partly based on a subjective impression of the patient. As there is no record of this implicit assessment of fitness, for the purpose of research this confounder is unobserved. Only the patient characteristics like performance score and age are available.

In addition to the overall fitness of a patient and the treatment they receive, an important factor for the variation in patient outcomes is the biological behavior of the tumor. There are different biomarkers of tumor behavior that are known in the clinical process, like histologic subtype or intra-tumor genetic heterogeneity *(8)*. These biomarkers are related to prognosis and / or the treatment effect. Tumor behavior is fundamentally unobservable in the sense that neither the physician nor the researcher observe this behavior fully. Only the biomarkers are available.

Patient fitness and tumor behavior are thus two unobserved variables that induce relationships between the observed patient characteristics, the biomarkers, the treatment decision and the outcome. The causal relationships between these variables are represented in a causal directed acyclic graph (DAG), shown in Figure *1*. For different cancer settings, different markers of patient fitness and tumor behavior may be relevant. Filling in application specific variables in the DAG is the first step of the PROTECT method. It could be that multiple choices for sets of variables are possible for a specific application. As explained in the appendix (section methods, estimation in a marginalized DAG), the PROTECT average treatment effect estimate is insensitive to the specific choice of variables.

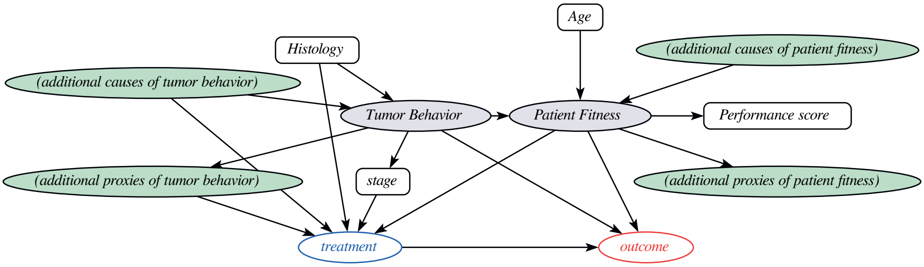

Age
Histology
(additional causes of patient fitness)
(additional causes of tumor behavior)
Tumor Behavior
Patient Fitness
Performance score
(additional proxies of tumor behavior)
stage
(additional proxies of patient fitness)
treatment
outcome

Figure 1. The behavior-fitness causal Directed Acyclic Graph (DAG) scaffold for cancer treatment decisions. Circles indicate variables, grey-shaded variables are unobserved. Arrows point from a cause variable to an effect variable. Tumor behavior and patient fitness are unobserved variables that induce correlations between the observed variables. The definition of the treatment variable and potentially the outcome variable vary per cancer setting. Depending on the specific situation, relevant additional cause variables and effect variables for tumor behavior and patient fitness should be selected. Estimating the effect of the treatment on the outcome (potentially conditional on the other variables in the DAG) is the target application of PROTECT. The presence of the unobserved confounder fitness implies that conventional confounding adjustment methods cannot be used to estimate treatment effects from observational data, whereas the proposed method PROTECT can. Filling in additional proxies and causes of tumor behavior and patient fitness in this DAG is the first step of PROTECT. PROTECT: proxy based individual treatment effect modeling in cancer.

From the DAG to treatment effect estimation

Having established a DAG for observational cancer research, the question is if and how the treatment effect can be estimated from observational data. To answer this, we use Pearl's Structural Causal Models framework *(9)*. The presence of the unobserved confounder overall fitness implies estimates of the treatment effect by direct conditioning on the observed variables (e.g. through multivariable regression or propensity score-based reweighting) would be incorrect as the back-door criterion is not fulfilled *(9–13)*. This does not rule out the possibility to estimate the treatment effect. Several methods exist for estimating treatment effects when there are proxy variables of unobserved confounders. Proxy variables are variables that are caused by the confounder, but do not causally influence the treatment decision and the outcome. Performance score is an example of a proxy variable for the confounder overall fitness, as performance score depends on fitness but does not cause overall survival or the treatment decision directly. One class of proxy methods relies on bridge functions *(12, 14–17)*. These methods leverage information from proxy measurements to reconstruct a bridge function that is sufficient for treatment effect estimation. To know if such a function can be estimated from the observed variables, these methods require additional assumptions. A frequently required assumption is that all variables are discrete *(12, 14, 17)*, which is an unnatural assumption for the confounder overall fitness, or that all variables are continuous *(12, 14)*, which is rarely the case in medical research. More flexible bridge function methods exist but these methods require complicated estimation procedures that require large sample sizes, which makes these methods unsuitable for many medical applications *(15)*.

An alternative approach is by estimating the joint distribution of the observed variables and the unobserved variables, using only the observed variables *(18)*. When sufficient information on the data generating process is available, this joint distribution can be estimated by modeling the data generative process directly. With this approach, each variable in the DAG is associated with an explicit structural equation that depends on the direct cause variables of this variable and random noise. If the joint distribution can be estimated, the treatment effect can still be calculated because the back-door adjustment formula can be applied using the estimated distributions of the outcome given treatment and fitness, and fitness given the cause variables and proxy variables of fitness *(9, 18)*.

Both proxy-based approaches require assumptions in addition to the DAG. These assumptions should be based on background knowledge. Background knowledge naturally comes in the form of parts of the data generating process.

Modeling the data generating process directly thus makes formulating the right assumptions easier for clinicians and researchers. Moreover, it makes it more accessible for readers to assess the validity of the made assumptions. One example of such an assumption is the statement that performance score should be better for patients with higher overall fitness. This assumption can be expressed as a monotonicity restriction in the structural equation for performance score and helps estimating the joint distribution of observed and unobserved variables and thus the treatment effect. In PROTECT, the joint distribution is estimated by specifying parametric forms for all the structural equations. If the parameters for the structural equations can be uniquely estimated from the observed data, the treatment effect can be estimated despite the unobserved confounder (see also appendix, section methods, treatment effect estimation). Translating background knowledge into specified parametric distributions for the observed and unobserved variables in the DAG thus forms the second step of PROTECT.

Model selection

It is possible that multiple choices for distributions are compatible with the available background knowledge for a specific research question. To reduce dependence of the treatment effect estimate on the specific choices for distributions, we introduce a data-driven model selection procedure. The model selection procedure is motivated by the fact that in the DAG, the unobserved variable overall fitness induces correlations between the proxy variables, treatment and outcome. The procedure uses cross-validation to verify whether these correlations are present in an estimated model by comparing the model predictions for one of the effect variables of fitness (proxies, treatment and outcome) with regression models based on only the direct causes of this variable, excluding fitness. If multiple models are selected in this step, they can be combined using a Bayesian model average. The appendix (section methods, model selection) contains a detailed report of this model selection procedure. After selection and estimation of the final model, the individual treatment effect is calculated as the difference in the expected outcome under the different treatments for each patient, conditional on their observed pre-treatment characteristics.

Computation

Once the parametric distributions are fully specified, the posterior distribution over the parameters for the structural equations can be estimated using Markov chain Monte Carlo (MCMC) sampling. We implemented PROTECT using state-of-the art inference techniques. Specifically we employ the No-U-Turn Hamiltonian Monte Carlo *(19)* implementation from the NumPyro package *(20)*, as NumPyro has JAX *(21)* as a back-end, enabling parallelized GPU-accelerated MCMC sampling. The code that implements PROTECT will be made freely available at a public online repository.

Application to stage III NSCLC

We now apply PROTECT to stage III NSCLC patients to estimate the relative effect of concurrent chemoradiation versus sequential chemoradiation on overall survival measured from the day of the treatment decision.

PROTECT step 1: definition of proxy variables and cause variables

In multiple discussion rounds with experts in pulmonary oncology, radiation oncology and radiology, the relevant proxy variables and cause variables for stage III NSCLC were selected. The additional variables included in the stage III NSCLC DAG are weight loss and estimated glomerular filtration rate (eGFR). Weight loss is an important proxy of tumor behavior, as patients with aggressive tumors tend to lose more weight due to the high disease burden. Renal function is an additional proxy variable for overall fitness in stage III NSCLC. The DAG is presented in Figure 2.

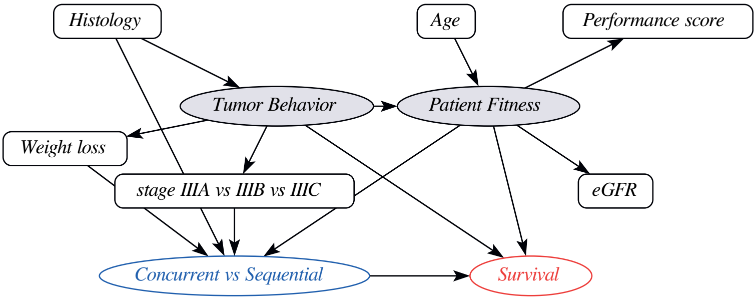


Figure 2. Causal Directed Acyclic Graph (DAG) with the variables involved in the treatment selection process and overall

survival for stage III non-small cell lung cancer patients. Circles indicate variables, shaded variables are unobserved. Arrows point from a cause variable to an effect variable. This DAG is a direct extension of the behavior-fitness DAG scaffold from the PROTECT method in Figure 1. PROTECT: proxy based individual treatment effect modeling in cancer; eGFR: estimated glomerular filtration rate.

PROTECT step 2 and 3: specification of distributions and model selection

We parameterized the joint distribution of observed variables and unobserved variables by specifying linear models for all structural equations indicated by the DAG, with link functions and error distributions as appropriate (i.e. linear regression for continuous variables, logistic regression for binary variables, a linear proportional hazards model for survival *(22))*. To estimate individual treatment effects, the model for survival was augmented with treatment – covariate interaction terms. Details on the exact formulation of the model variants entered in the model selection procedure are presented in the appendix (section methods, pre-processing, parametric models and priors).

Study Population

We identified 844 patients with 864 episodes of stage III NSCLC in 9 hospitals in the Utrecht region, the Netherlands, treated between 2009 and 2018. A total of 360 episodes were excluded based on the following exclusion criteria (more than one criterion can apply to an episode): the primary treatment plan had palliative intent (N=140), the primary treatment plan included surgery (N=49), the presence of a concurrent other tumor, including a second NSCLC (N=35), local re-irradiation for the recurrence of an earlier episode (N=9), Pancoast tumor (N=9), having a second episode of stage III NSCLC (N=8, only the second episode was excluded from the analysis), receiving radiotherapy in a different hospital (N=3), chemotherapy and radiotherapy in reversed order to prevent spinal canal invasion (N=3), emigration (N=1). Finally, 123 patients were excluded due to a missing value for weight loss.

The mean age of the 504 included patients was 64.9 (range 37 - 86), of which 300 (59.5%) were male. Substage IIIA accounted for 268 of the cases (53%). We observed 141 deaths in 224 patients who underwent concurrent chemoradiation (632 patient years) and 214 deaths in 280 patients who underwent sequential chemoradiation (603 patient years). Compared with the study population from the meta-analysis of RCTs *(6)*, our patients were older (median, 61.7 vs 66) and had worse performance scores (ECOG of 2 or greater: 1% vs 10%). In the appendix (section results, supplemental tables) a table with an extensive comparison is presented. The start of follow-up was imputed for 12.3% of the patients. The median survival time was 1.87 years, the median follow-up time for patients who were censored was 3.80 years. The last date of follow-up was February 6th, 2020. Patient characteristics are summarized in Table 1.

Table 1. Baseline characteristics stratified by chemoradiation type: concurrent chemoradiation or sequential chemoradiation.

	concurrent	**sequential**
n	224	280
age (mean (SD))	61.21 (9.25)	67.83 (8.95)
substage (%)		
IIIA	141 (62.9)	127 (45.4)
IIIB	81 (36.2)	151 (53.9)
IIIC	0 (0.0)	2 (0.7)
missing	2 (0.9)	0 (0.0)
weight loss > 3% (%)	99 (44.2)	127 (45.4)
ECOG PS (%)		
0	139 (62.1)	119 (42.5)
1	67 (29.9)	107 (38.2)
2	6 (2.7)	43 (15.4)
3	1 (0.4)	1 (0.4)
missing	11 (4.9)	10 (3.6)
eGFR (%)		
<60 ml / min / 1.73m^2	11 (4.9)	26 (9.3)
>=60 ml / min / 1.73m^2	175 (78.1)	163 (58.2)
missing	38 (17.0)	91 (32.5)
histology (%)		

Table 1. Continued.

	concurrent	sequential
adeno carcinoma	107 (47.8)	93 (33.2)
squamous cell carcinoma	61 (27.2)	133 (47.5)
other	47 (21.0)	42 (15.0)
missing	9 (4.0)	12 (4.3)
deceased during follow-up (%)	141 (62.9)	214 (76.4)
male sex (%)	130 (58.0)	170 (60.7)

Weight loss is defined as weight loss over 3% of the original weight over the six months preceding the start of follow-up. ECOG PS: Eastern Cooperative Oncology Group performance score; SD: standard deviation; eGFR: estimated glomerular filtration rate.

Treatment Effect Estimation

Overall survival was significantly better for patients with concurrent chemoradiation compared to sequential chemoradiation (hazard ratio, 0.66; 95% confidence interval, 0.53 to 0.82). When estimating the treatment effect with multivariable Cox-regression, adjusting for age, histology, weight loss, clinical substage, performance score and eGFR, concurrent treatment had a favorable survival (hazard ratio, 0.81; 95% confidence interval, 0.60 to 1.09). This treatment effect estimate is more extreme than the effect reported in the meta-analysis of RCTs (6) and possibly affected by residual confounding bias. In contrast, the average treatment effect estimated using PROTECT showed no benefit of concurrent over sequential treatment on average in our population (hazard ratio, 1.01; 95% credible interval, 0.68 to 1.53). An overview of the treatment effects is presented in Figure 3.

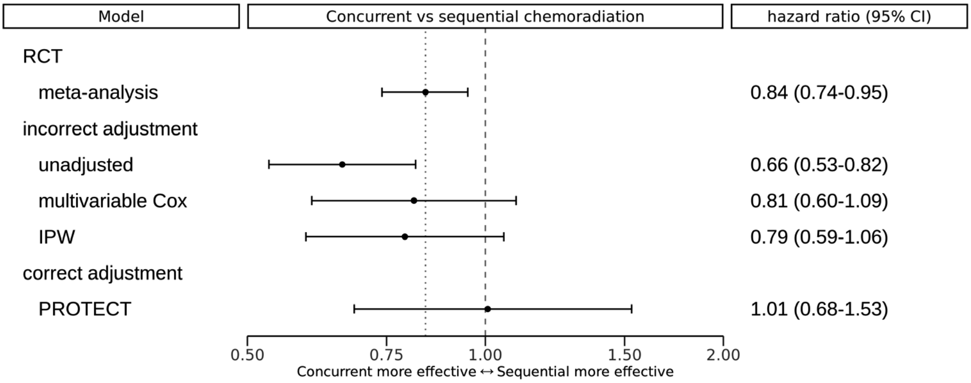


Figure 3. Overview of treatment effects estimated with different methods. The dashed vertical reference line indicates the null effect (hazard ratio of 1), the dotted reference line indicates the point estimate of the meta-analysis of RCTs by Aupérin et al. (6). IPW: inverse-probability of treatment weighted Cox-proportional hazards model. PROTECT: proxy based individual treatment effect modeling in cancer; CI: confidence interval, for PROTECT: credible interval; RCT: randomized controlled trial.

Treatment Effect Modification and Individual Treatment Recommendations

According to the PROTECT individual treatment effect model estimate, the following variables were associated with a reduced effectiveness of concurrent treatment: clinical substage IIIB and IIIC, the presence of weight loss and adenocarcinoma histologic subtype. Age, ECOG performance score and eGFR were not related to treatment efficacy. Treatment effect modifications per variable are presented in Figure 4.

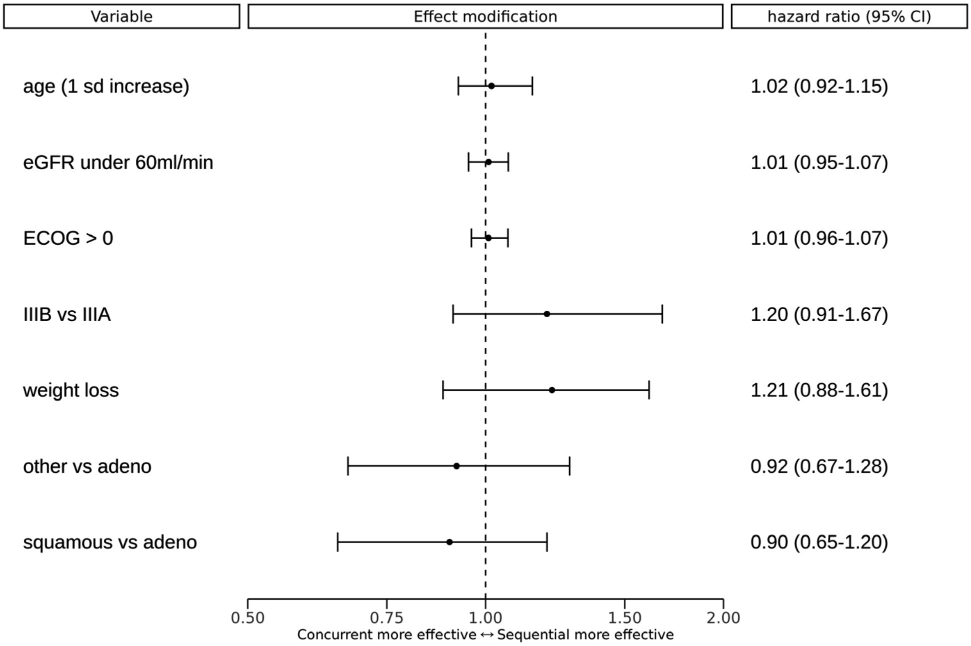


Figure 4. Differences in estimated treatment effect compared to the average treatment effect for a one unit increase per variable. A unit increase means switching from 'no' to 'yes' for binary variables, and a 1 standard deviation increase from the mean for continuous variables (age). These are step-function versions of the partial dependence functions as described by Friedman (23). 'other vs adeno' indicates the effect modification of other histology type compared to adenocarcinoma. 'squamous vs adeno' indicates the effect modification of squamous cell carcinoma compared to adenocarcinoma. CI: credible interval, eGFR: estimated glomerular filtration rate, ECOG: Eastern Cooperative Oncology Group performance score, IIIB: clinical stage IIIB or IIIC, IIIA: clinical stage IIIA, Weight loss is defined as weight loss over 3% of the original weight over the six months preceding the start of follow-up.

For each patient PROTECT predicted the probability that concurrent treatment would lead to improved expected survival compared with sequential treatment, based on the pre-treatment variables. For 274 out of 504 patients (54.4%) this probability was greater than 50%. See Figure 5 for an overview of the predicted treatment benefit expressed as a hazard ratio per patient.

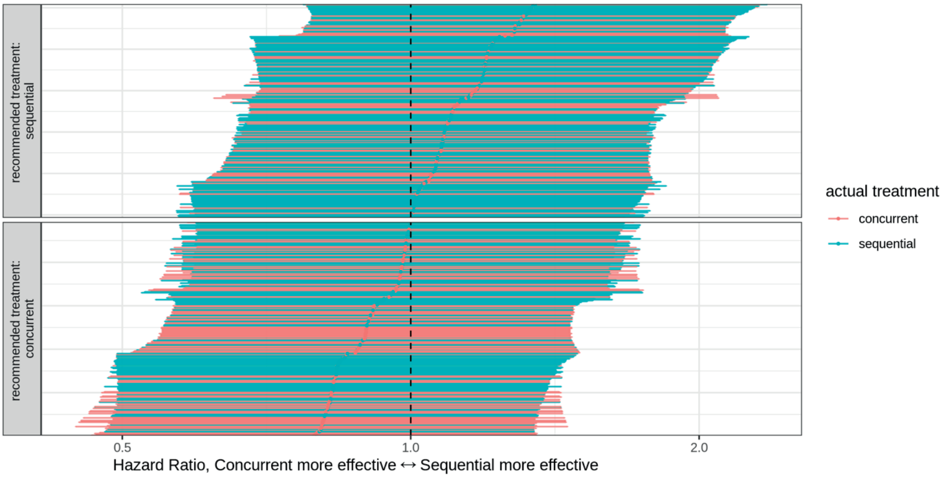


Figure 5. Predicted individual treatment effects for all 504 included patients. Each patient is represented by a horizontal line indicating the 95% credible interval of the predicted hazard ratio for overall survival of concurrent chemoradiation versus sequential chemoradiation, and the point estimate. Colors are added to indicate the actually received treatment. The reference line indicates the null effect: both treatments are equally effective.

As an additional model check, we investigated whether patients with lower estimated overall fitness were more likely to discontinue treatment due to side-effects or disease progression. After estimating patient fitness based on pre-treatment variables, we observed that in patients who underwent concurrent chemoradiation, a lower estimated fitness was associated with a higher probability of discontinuing the original treatment plan after the initiation of treatment (Area Under the Curve of the Receiver Operating Characteristic (AUC), 0.61; 95% confidence interval, 0.55 to 0.68).

Sampling diagnostics and model fit

We assessed the convergence of the MCMC chains and inspected the possible existence of multiple posterior modes that would prohibit the identification of the joint distribution and thus the treatment effect. The maximum Gelman-Rubin r-hat statistic across all parameters was 1.002 and the posterior density plots were unimodal. Given the large number of independent chains (16) and the number of samples per chain (7500 samples following 2500 warm-up samples) it is unlikely that other posterior modes exist. As outlined earlier, a unimodal posterior distribution implies identification of the treatment effect given the observed data and the modeling assumptions, despite the unobserved confounder. There were 231 divergent transitions in 120,000 samples (0.2%), after inspection of parameter trace plots these were deemed false positives.

The concordance index for overall survival predictions based on pre-treatment variables and the given treatment was 0.59 (95% confidence interval, 0.56 to 0.62). The AUC for predicting the treatment decision based on pre-treatment variables was 0.76 (95% confidence interval, 0.72-0.80).

Sensitivity analyses

We performed three sensitivity analyses of the average treatment effect estimate. The appendix (section sensitivity analyses) contains additional details on the justification, methods and results of these analyses.

First, we assessed the sensitivity of the average treatment effect estimate to an unmodeled confounder by re-estimating the model with an additional unobserved confounder with a known relationship with treatment and survival. We used several combinations of values for the association between this extra unobserved confounder and treatment and survival. We found that this confounder had to have a confounding strength that was more than half as strong as the modeled unobserved fitness, but opposite in sign with respect to survival, to drive the point estimate of the average treatment effect to a value more extreme than the treatment effect estimate from the RCTs.

Secondly, we assessed the potential bias induced by non-random missing values in weight loss. It is possible that weight loss is more often recorded in the EHR when pronounced weight loss is present. For this specific missingness pattern, treatment effect estimation using the complete cases is unbiased, while imputation may lead to bias *(24)*. This is why we excluded patients with a missing value of weight loss. As weight loss may be related to treatment efficacy and the true prevalence of weight loss is unknown due to the missing data, we recalculated the average treatment effect for several hypothetical values of the prevalence of weight loss by reweighting patients according to their weight loss. In the most extreme hypothetical case where weight loss was always observed if it was present, the average treatment effect estimate (hazard ratio, 1.01; 95% credible interval, 0.65 to 1.53) was very close to the other extreme case where weight loss was missing completely at random (hazard ratio, 1.03; 95% credible interval, 0.71 to 1.68).

Finally, we calculated what the estimated average treatment effect would be when restricting the analysis to a subsample of the cohort that is more like the population of the RCTs *(6)*. Under the assumption that the mechanism for selection for concurrent treatment and the mechanism for selection for inclusion in the RCT are similar, the population was restricted to those with a predicted probability of concurrent treatment higher than several different cut-offs (0%, 25%, 50%, 75%). When restricting the analysis to patients with a higher predicted probability of concurrent treatment to approximate the RCT population, the estimated treatment effect shifted towards concurrent chemoradiation being more effective (hazard ratio 1.01 for predicted probability >0%, N=504; hazard ratio, 1.00 for >25%, N=359; hazard ratio 0.98 for >50%, N=193; hazard ratio 0.95 for >75%, N=55). The treatment effect moves in the direction of the RCT estimate but the estimated survival benefit of concurrent treatment is still smaller. It may be that the method used to match our population with the RCT population was too crude.

Discussion

We present PROTECT, a method that uses proxy measurements of unobserved confounders to estimate treatment effects from observational data. PROTECT addresses a pervasive problem in observational cancer research: the lack of a direct measurement of the confounder overall fitness. When applied to stage III NSCLC, the results indicate that on average the reported benefit of concurrent over sequential chemoradiation for overall survival may be absent in a real-world population with more patients in lower overall fitness. In our cohort just over half of the patients were treated with sequential chemoradiation, which would imply that in approximately half of the patients the treating physician was confident that concurrent treatment would be beneficial. This statistic fits in with the absence of an average treatment effect. Whereas conventional confounding adjustment methods find a more extreme treatment effect than reported in RCTs, the results from PROTECT are in line with the recommendations from guidelines that patients with lower overall fitness are less likely to benefit from concurrent treatment. This positive association between overall fitness and treatment effect could be due to a higher risk of treatment discontinuation among patients with lower overall fitness when they are assigned the concurrent chemoradiation treatment regimen. Even though the model was not directly optimized to predict discontinuation of treatment, patients for whom the model estimated lower fitness where indeed more likely to discontinue treatment if they were assigned to concurrent treatment. In contrast, the meta-analysis by Aupérin et al. did not find a statistically significant treatment effect modification by age or ECOG performance score *(6)*. This could be due to the patient inclusion mechanism. All included patients were deemed fit enough for concurrent treatment. This means that older patients included in the RCTs are likely to have been relatively fit for their age. The same principle holds for performance score. If the treatment effect is indeed modified by overall fitness as our results and treatment guidelines suggest, it could be that the variation in fitness is too low in the RCT population to detect the treatment effect modification.

As this is a non-randomized study it is impossible to rule out confounding bias. Several steps were taken to mitigate potential confounding bias. First, we identified potential confounders from literature and domain expertise. We then applied a data-driven model selection procedure that rejects models that do not conform with the confounding structure implicated by the DAG. Lastly, a sensitivity analysis with an independent omitted confounder showed that the results are robust to unobserved confounders of reasonable strength.

Due to the moderate study sample size, our study does not attain high precision in treatment effect estimation. To address this, future studies should be based on data from larger consortia. Furthermore, the discriminatory power of our model for overall survival was low. This could be due to the omission of other important prognostic biomarkers in the analysis, or due to the intrinsic randomness in overall survival time for cancer patients. Our concordance index is in line with a recent meta-analysis of prognostic models for NSCLC patients treated with curative radiotherapy *(25)*.

Most of our patients were treated before approval of durvalumab for stage III NSCLC in the Netherlands. As treatment with durvalumab is contingent on successfully completing the chemotherapy and radiotherapy, the presented model is still of use as the predictions are correlated with successful completion of the treatment regimen. Since durvalumab improves overall survival *(26)*, successfully completing the treatment may become relatively more important than whether the initial treatment was concurrent or sequential chemoradiation.

RCTs remain crucial for treatment effect estimation for cancer patients as they do not suffer from confounding bias. Still there are several situations where treatment effect estimates from observational data are desirable. When parts of the real-world population are not covered by the RCTs for a certain treatment but observational data is available, PROTECT can be used to estimate the treatment effect in these subpopulations. Furthermore, as the method can estimate both average treatment effects and individual treatment effects, PROTECT can be used for studies on biomarkers of treatment efficacy. When new biomarkers become available that were not measured in RCTs, the treatment effect modification of this biomarker can be studied in an observational cohort using PROTECT. In both applications, the resulting estimates may indicate that a new RCT is warranted in specific subpopulations. In this way, observational studies may supplement evidence from RCTs. Conversely, RCTs provide a point of reference for observational studies.

To facilitate future applications of PROTECT, a three-step overview of PROTECT is presented in Table 2. In the appendix (section discussion) we present two examples where PROTECT could be applied, one in unresectable laryngeal carcinoma and one in stage III squamous cell esophageal cancer. In each application there may be additional confounders to consider. However, the core of the PROTECT DAG will be applicable to many different cancer types.

Table 2. Overview of the PROTECT method for developing individual treatment effect models from observational cancer cohorts.

Step 1	Specify proxies and causes in the behavior-fitness DAG scaffold
Step 2	Specify parametric distributions for observed and unobserved variables
Step 3	Apply PROTECT model selection criteria

Notes per step: 1. The causal Directed Acyclic Graph (DAG) describes the causal relationships between variables involved in the treatment decisions and outcomes. See Figure 1 for the behavior-fitness DAG scaffold for observational cancer research. To apply PROTECT, researchers need to specify the definitions of treatment, potential additional proxies and causes of tumor behavior and patient fitness, and possibly the outcome. Potential additional application-specific sources of confounding or selection bias must be added as well. 2. These choices should be based on background knowledge. Specific care should be taken to check whether these distributions are uniquely identified. This depends on both the number of proxies of fitness (more is better) and the flexibility of the statistical models. 3. As multiple choices can be made in step 2, applying the PROTECT model selection criteria will reduce the dependence of treatment effect estimates on parameterization choices. Models that do not conform to the confounding structure in the DAG will be rejected. DAG: causal Directed Acyclic Graph; PROTECT: Proxy based individual treatment effect modeling in cancer

In conclusion we present PROTECT, a method for individual treatment effect estimation for cancer patients in the presence of unobserved confounders using proxy measurements. When applied to a real-world stage III NSCLC cohort, PROTECT provided credible treatment effect estimates whereas conventional confounding adjustment methods did not.

Materials and Methods

Study design

Data Source

We conducted a retrospective observational cohort study at the Department of Radiotherapy of the University Medical Center Utrecht, the Netherlands. Patients were referred to the Utrecht center for radiotherapy from the thoracic oncology departments of 9 different hospitals in the Utrecht region in the Netherlands. This study was conducted in accordance with the applicable privacy regulations. As this was a non-experimental retrospective study and most of the patients had died, a waiver for informed consent was obtained from the institutional review board at the University Medical Center Utrecht, along with approval of the study protocol (protocol number WAG/dgv/18/005984). All the methods were performed in accordance with the Declaration of Helsinki.

Cohort Selection

Patients referred to our center for the consideration of curative (chemo) radiotherapy as a primary therapy for a first episode of clinical stage III NSCLC between November 2009 and December 2018 were retrospectively identified. Patients had been staged according to the American Joint Committee on Cancer Tumor-Node-Metastasis (TNM) staging protocol. We maintained the TNM version that was clinically used at the time of treatment, spanning versions six *(27)*, seven *(28)* and eight *(29)*. We excluded patients who were eligible for primary surgery or who were treated with palliative intent. Other exclusion criteria were a concurrent other tumor (including a second NSCLC), a prior diagnosis of stage III NSCLC, receiving radiotherapy before chemotherapy to prevent spinal canal invasion, having a Pancoast tumor, receiving radiotherapy at another institution or emigration during follow-up. Patients who were seen at the radiotherapy outpatient clinic but for some reason did not receive radiotherapy were not excluded from the analysis as they are part of the target population for the individual treatment effect model.

Definition of intervention and outcome

Clinical variables were extracted from the electronic health records (EHR) which includes referral letters for patients from other hospitals that were referred to our hospital for radiotherapy. Concurrent chemoradiation was defined as a combination of chemotherapy and radiotherapy with time overlap between the treatments, whereas for sequential chemoradiation the start of radiotherapy was subsequent to administration of the last chemotherapy cycle. For both treatments, the planned radiotherapy had to consist of a definitive physical dose of 54 Gray or higher *(30)*. Chemotherapy consisted of two to four cycles of platinum-based chemotherapy (cisplatin or carboplatin, with etoposide, gemcitabine or pemetrexed). As this is a multi-institutional non-experimental study, chemotherapy regimens varied. The goal of an individual treatment effect model is to influence future treatment decisions. Therefore, the intervention under study was concurrent versus sequential chemoradiation according to the initial treatment decision. This choice is in line with the general preference for intention-to-treat analyses in RCTs *(31)*.

The start of follow-up was defined as the date of the last multi-disciplinary tumor board meeting preceding the start of treatment, as this is generally the moment the treatment decision is made. Specific care was taken to record the values of variables as they were known at this time point. The outcome was overall survival measured on a continuous time scale. If no date of death was noted in the EHR, data for overall survival was supplemented by querying the Dutch Personal Records Database.

Statistical Analysis

Covariates

The set of variables for the analysis consisted of age, histology (grouped as adenocarcinoma, squamous cell carcinoma, or other), the presence of any weight loss (defined as > 3% of original weight in the 6 months leading up to the treatment decision), performance status 0 versus 1 or higher, defined by the ECOG standard *(32)*, eGFR higher or lower than 60 ml / minute / 1.73 m^2 and TNM stage IIIA vs IIIB or IIIC.

Missing Data Handling

Variables with less than 5% missing values were imputed using mean imputation for continuous variables or a fixed value of 0.5 for binary variables. When the date of the tumor board was unknown this date was imputed based on the date of treatment start with mean imputation per treatment category. Missing values in proxy variables of fitness are assumed to be missing at random conditionally on the observed variables. Further, we assumed that the missingness of weight loss was dependent on the presence of weight loss. For this specific missingness pattern, complete case analysis is unbiased, while imputation may lead to bias *(24)*. Therefore, we excluded patients with a missing value of weight loss. In the appendix (section methods, missing data) we elaborate on this assumption further and present a sensitivity analysis regarding the missingness in weight loss (section sensitivity analyses).

Model Evaluation

The treatment effect estimates from PROTECT were contrasted with the baseline approach of including the observed variables in a multivariable Cox-proportional hazards model and an inverse propensity score weighted Cox-proportional hazards model. Estimated treatment effects were compared with the reference value from the meta-analysis by Aupérin *(6)*. As described in the appendix (section methods, pre-processing, parametric models and priors) the model includes a non-linear component. Therefore, potential treatment effect modification for a variable was inspected using partial dependence functions *(23)*.

Model fit for overall survival was assessed using Harrell's concordance index *(33)*. Model fit for the treatment choice was assessed with the AUC.

Posterior samples were simulated using 16 independently initialized MCMC chains with 7500 samples each, following 2500 warm-up samples. The mixing of chains was inspected with the Gelman-Rubin r-hat statistic *(34)* and the presence of multiple posterior modes was checked visually from posterior density plots.

As an additional model evaluation, we estimated the association between the estimated fitness based on pre-treatment variables and the occurrence of a negative treatment switch anywhere during the treatment. A negative treatment switch was defined as any reduction in treatment intensity compared to the original treatment intention, occurring after the first day of treatment. This included a reduction in chemotherapy dose, fewer chemotherapy cycles, a switch from concurrent to sequential chemoradiation, a lower radiotherapy dose or complete cessation of treatment.

Sensitivity analyses

We tested the robustness of the average treatment effect estimate to a hypothetical omitted confounder. This was done by re-estimating the model with an additional unobserved variable with several hypothetical relationships with the treatment and the outcome. Finally, we calculated what the estimated average treatment effect would be when restricting the analysis to a subsample of the cohort that is more like the population of the RCTs *(6)*. Under the assumption that the mechanism for selection for concurrent treatment and the mechanism for selection for inclusion in the RCT are similar, the population was restricted to those with a predicted probability of concurrent treatment higher than several different cut-offs (0%, 25%, 50%, 75%). Details on the justification and implementation of these sensitivity analyses are presented in the appendix (section sensitivity analyses).

Implementation

NumPyro version 0.4.1 and JAX version 0.2.7 were used for model estimation. R version 4.0.3 was used for model evaluations.

Reporting

For reporting, we adhered to the STROBE reporting guidelines for observational research *(35)*. A completed form is available in the supplemental material.

Data availability

Due to local privacy regulations, the original patient data cannot be shared. The code that implements the statistical models and model selection procedure will be made publicly available at an online repository.

References

[1] C. M. Booth, I. F. Tannock, Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence, *British Journal of Cancer* **110**, 551–555 (2014).

[2] J. H. Lewis, M. L. Kilgore, D. P. Goldman, E. L. Trimble, R. Kaplan, M. J. Montello, M. G. Housman, J. J. Escarce, Participation of Patients 65 Years of Age or Older in Cancer Clinical Trials, *JCO* **21**, 1383–1389 (2003).

[3] S. K. Vinod, Decision making in lung cancer – how applicable are the guidelines?, *Clin Oncol (R Coll Radiol)* **27**, 125–131 (2015).

[4] FDA-NIH Biomarker Working Group, *Predictive Biomarker* (Food and Drug Administration (US), 2016; https://www.ncbi.nlm.nih.gov/books/NBK402283/).

[5] D. S. Ettinger, NCCN Non-Small Cell Lung Cancer Guideline, Version 1.2021 (2020) (available at https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf).

[6] A. Aupérin, C. Le Péchoux, E. Rolland, W. J. Curran, K. Furuse, P. Fournel, J. Belderbos, G. Clamon, H. C. Ulutin, R. Paulus, T. Yamanaka, M.-C. Bozonnat, A. Uitterhoeve, X. Wang, L. Stewart, R. Arriagada, S. Burdett, J.-P. Pignon, Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non-small-cell lung cancer, *J. Clin. Oncol.* **28**, 2181–2190 (2010).

[7] N. Ramnath, T. J. Dilling, L. J. Harris, A. W. Kim, G. C. Michaud, A. A. Balekian, R. Diekemper, F. C. Detterbeck, D. A. Arenberg, Treatment of Stage III Non-small Cell Lung Cancer, *Chest* **143**, e314S-e340S (2013).

[8] M. Jamal-Hanjani, G. A. Wilson, N. McGranahan, N. J. Birkbak, T. B. K. Watkins, S. Veeriah, S. Shafi, D. H. Johnson, R. Mitter, R. Rosenthal, M. Salm, S. Horswell, M. Escudero, N. Matthews, A. Rowan, T. Chambers, D. A. Moore, S. Turajlic, H. Xu, S.-M. Lee, M. D. Forster, T. Ahmad, C. T. Hiley, C. Abbosh, M. Falzon, E. Borg, T. Marafioti, D. Lawrence, M. Hayward, S. Kolvekar, N. Panagiotopoulos, S. M. Janes, R. Thakrar, A. Ahmed, F. Blackhall, Y. Summers, R. Shah, L. Joseph, A. M. Quinn, P. A. Crosbie, B. Naidu, G. Middleton, G. Langman, S. Trotter, M. Nicolson, H. Remmen, K. Kerr, M. Chetty, L. Gomersall, D. A. Fennell, A. Nakas, S. Rathinam, G. Anand, S. Khan, P. Russell, V. Ezhil, B. Ismail, M. Irvin-Sellers, V. Prakash, J. F. Lester, M. Kornaszewska, R. Attanoos, H. Adams, H. Davies, S. Dentro, P. Taniere, B. O'Sullivan, H. L. Lowe, J. A. Hartley, N. Iles, H. Bell, Y. Ngai, J. A. Shaw, J. Herrero, Z. Szallasi, R. F. Schwarz, A. Stewart, S. A. Quezada, J. Le Quesne, P. Van Loo, C. Dive, A. Hackshaw, C. Swanton, Tracking the Evolution of Non–Small-Cell Lung Cancer, *New England Journal of Medicine* **376**, 2109–2121 (2017).

[9] J. Pearl, Ed., in *Causality*, (Cambridge University Press, Cambridge, 2009), pp. 65–106.

[10] S. Greenland, The Effect of Misclassification in the Presence of Covariates, *American Journal of Epidemiology* **112**, 564–569 (1980).

[11] E. Ogburn, T. Vanderweele, Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders, *Biometrika* **100**, 241–248 (2013).

[12] M. Kuroki, J. Pearl, Measurement bias and effect restoration in causal inference, *Biometrika* **101**, 423–437 (2014).

[13] P. R. ROSENBAUM, D. B. RUBIN, The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55 (1983).

[14] W. Miao, Z. Geng, E. T. Tchetgen, Identifying Causal Effects With Proxy Variables of an Unmeasured Confounder, (2016).

[15] N. Kallus, X. Mao, M. Uehara, Causal Inference Under Unmeasured Confounding With Negative Controls: A Minimax Learning Approach, *arXiv:2103.14029 [cs, stat]* (2021) (available at http://arxiv.org/abs/2103.14029).

[16] Y. Wang, D. M. Blei, The Blessings of Multiple Causes, *Journal of the American Statistical Association* **114**, 1574–1596 (2019).

[17] S. Lee, E. Bareinboim, Causal Identification with Matrix Equations, *Columbia CausalAI Laboratory Technical Report (R-70)* (2021).

[18] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, M. Welling, Causal Effect Inference with Deep Latent-Variable Models, (2017).

[19] M. D. Hoffman, A. Gelman, The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, *arXiv:1111.4246 [cs, stat]* (2011) (available at http://arxiv.org/abs/1111.4246).

[20] D. Phan, N. Pradhan, M. Jankowiak, Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro, *arXiv preprint arXiv:1912.11554* (2019).

[21] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, Q. Zhang, *JAX: composable transformations of Python+NumPy programs* (2018; http://github.com/google/jax).

[22] K. Burke, M. C. Jones, A. Noufaily, A Flexible Parametric Modelling Framework for Survival Analysis, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **69**, 429–457 (2020).

[23] J. H. Friedman, Greedy function approximation: A gradient boosting machine., *Ann. Statist.* **29**, 1189–1232 (2001).

[24] I. R. White, J. B. Carlin, Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values, *Statistics in Medicine* **29**, 2920–2931 (2010).

[25] G. Kothari, J. Korte, E. J. Lehrer, N. G. Zaorsky, S. Lazarakis, T. Kron, N. Hardcastle, S. Siva, A systematic review and meta-analysis of the prognostic value of radiomics based models in non-small cell lung cancer treated with curative radiotherapy, *Radiotherapy and Oncology* **155**, 188–203 (2021).

[26] C. Faivre-Finn, D. Vicente, T. Kurata, D. Planchard, L. Paz-Ares, J. F. Vansteenkiste, D. R. Spigel, M. C. Garassino, M. Reck, S. Senan, J. Naidoo, A. Rimner, Y.-L. Wu, J. E. Gray, M. Özgüroğlu, K. H. Lee, B. C. Cho, T. Kato, M. de Wit, M. Newton, L. Wang, P. Thiyagarajah, S. J. Antonia, Four-Year Survival With Durvalumab After Chemoradiotherapy in Stage III NSCLC—an Update From the PACIFIC Trial, *Journal of Thoracic Oncology* **16**, 860–867 (2021).

[27] TNM Atlas, 6th Edition | Wiley*Wiley.com* (available at https://www.wiley.com/en-nl/TNM+Atlas%2C+6th+Edition-p-9781118695609).

[28] TNM Classification of Malignant Tumours, 7th Edition | Wiley*Wiley.com* (available at https://www.wiley.com/en-nl/TNM+Classification+of+Malignant+Tumours%2C+7th+Edition-p-9781444358964).

[29] TNM Classification of Malignant Tumours, 8th Edition | Wiley*Wiley.com* (available at https://www.wiley.com/en-us/TNM+Classification+of+Malignant+Tumours%2C+8th+Edition-p-9781119263579).

[30] S. J. Antonia, A. Villegas, D. Daniel, D. Vicente, S. Murakami, R. Hui, T. Yokoi, A. Chiappori, K. H. Lee, M. de Wit, B. C. Cho, M. Bourhaba, X. Quantin, T. Tokito, T. Mekhail, D. Planchard, Y.-C. Kim, C. S. Karapetis, S. Hiret, G. Ostoros, K. Kubota, J. E. Gray, L. Paz-Ares, J. de Castro Carpeño, C. Wadsworth, G. Melillo, H. Jiang, Y. Huang, P. A. Dennis, M. Özgüroğlu, PACIFIC Investigators, Durvalumab after Chemoradiotherapy in Stage III Non-Small-Cell Lung Cancer, *N. Engl. J. Med.* **377**, 1919–1929 (2017).

[31] S. K. Gupta, Intention-to-treat concept: A review, *Perspect Clin Res* **2**, 109–112 (2011).

[32] M. M. Oken, R. H. Creech, D. C. Tormey, J. Horton, T. E. Davis, E. T. McFadden, P. P. Carbone, Toxicity and response criteria of the Eastern Cooperative Oncology Group, *American Journal of Clinical Oncology* **5**, 649–656 (1982).

[33] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, R. A. Rosati, Evaluating the Yield of Medical Tests, *JAMA* **247**, 2543–2546 (1982).

[34] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, P.-C. Bürkner, Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC, *Bayesian Anal.* (2020), doi:10.1214/20-BA1221.

[35] E. von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, J. P. Vandenbroucke, STROBE Initiative, The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies, *Ann Intern Med* **147**, 573–577 (2007).

Supplemental Data:

A Methods

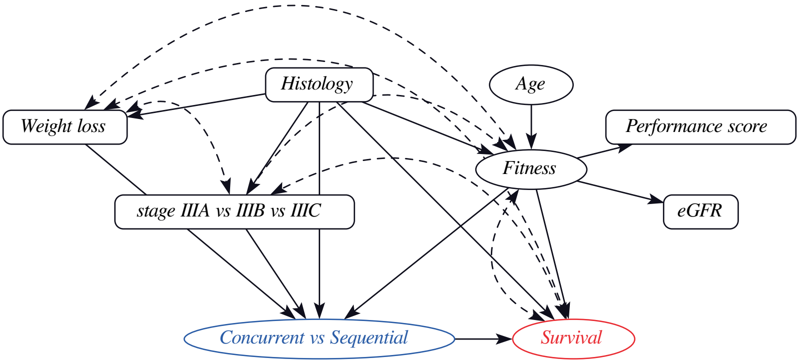


Figure 1: Acyclic Directed Mixed Graph, resulting from marginalizing out $TumorBehavior$ in the original DAG. A dashed bi-directed arrow indicates the presence of an unobserved common cause ($TumorBehavior$). eGFR: estimated Glomerular Filtration Rate

A.1 Treatment Effect Estimation with PROTECT

In this section we introduce additional implementation details of the PROTECT method.

A.1.1 Additional assumptions

In addition to the DAG, the PROTECT method relies on two other assumptions that are very common in treatment estimation from observational data. The first assumption is that of *overlap*. For the X-conditional average treatment effect to exist, when there are confounders L, the joint distributions of X, L of patients with different treatments should fully overlap [10]. The second assumption is the *Stable-Unit Treatment Value Assumption* [10]. This assumption has two components: *consistency* meaning that the observed factual outcome under a certain treatment is equal to the outcome that would be observed when intervening to make the same treatment decision; and *no interference*, meaning that assigning a treatment to a certain patient does not influence the interventional distribution of other patients.

A.1.2 Marginalizing out Tumor Behavior for Estimation

The PROTECT DAG has two latent factors: tumor behavior and patient fitness. Since the latent factors are not observed they need to be marginalized out, both during the model estimation phase and for posterior predictions on new data. As conditioning on the latent factor for tumor behavior is not needed to satisfy the backdoor rule, we can also do inference over the acyclic directed mixed graph (ADMG) [19] that results from marginalizing out tumor behavior, without harming the ability to estimate the conditional treatment effect. ADMGs are a generalization of DAGs that allow for bi-directed arrows, indicating the presence of an unobserved confounder [19]. The AMDG that results from marginalizing out tumor behavior has fewer parameters to estimate and was therefore used for model estimation. See Figure A.1.2 for the resulting ADMG. There are multiple bi-directed arrows in the ADMG as a result of marginalizing out tumor behavior. To model these dependencies, the original ancestral order is used. This means for instance that the outcome will be modeled conditional on stage and not vice-versa, as stage is an ancestor of the outcome in both the original DAG and the ADMG with tumor behavior marginalized out.

A.1.3 Target Estimand

Here we derive a non-parametric expression for estimating the model from the observed data. Let $t_x \in \{0, 1\}$ be the treatment indicator where. Let y denote the outcome. Let X and W be vectors of covariates (causes and proxies of fitness respectively, in line with the AMDG in Figure A.1.2). Let X^B denote the proxies and causes of tumor behavior from the original ADMG (histology, weight loss and stage IIIA vs IIIB vs IIIC in the case of stage III Non-Small Cell Lung Cancer). We use Pearl's do-operator to indicate intervening on a variable [17]. We are interested in the conditional average treatment effect (CATE) as the difference in expected survival under the different treatments:

$$\mathrm{CATE}(X, W, X^B) = \mathbf{E}\left[y|\mathrm{do}(t_x = 1), X, W, X^B\right] - \mathbf{E}\left[y|\mathrm{do}(t_x = 0), X, W, X^B\right] \tag{1}$$

The presented AMDG suggests a causal factorization of the conditional distribution in equation (1). We will now derive our target estimand of the joint distribution of the observable variables (y, t_x, W) given the control variables (X, X^B).

Proof. Let y be the outcome, t_x the treatment variable, $F \in \mathbb{R}$ a latent factor, W possibly multidimensional proxy variables for F, and X possibly multidimensional causes for F, X^B possibly multidimensional causes and proxies for tumor behavior, then

$$p(y, t_x, W|X, X^B)$$

$$=^a \int_F p(y, t_x, W|X, X^B, F)p(F|X, X^B)dF \tag{2}$$

$$=^b \int_F p(y|t_x, W, X, X^B, F)p(t_x|W, X, X^B, F)p(W|X, X^B, F)p(F|X, X^B)dF \tag{3}$$

$$=^c \int_F p(y|t_x, X^B, F)p(t_x|X^B, F)p(W|F)p(F|X, X^B)dF \tag{4}$$

a) by the law of total probability

b) by the chain rule of of probability

c) by conditional independencies implied by the ADMG

In 4 the outcome distribution conditions on all confounders, thereby satisfying the backdoor rule. This implies we can use Rule 2 of the rules of do-calculus and exchange observing t_x (as we do in the observational distribution from which our samples are drawn), with intervening on t_x (the interventional distribution that is the target of our research) [16].

After estimating the joint distribution, we can calculate the CATE for a patient by conditioning 4 on the observed proxy variables W for both $t_x = 0$ and $t_x = 1$ and marginalizing over F, and then calculating the difference in the expected value of y for $t_x = 1$ and $t_x = 0$. The average treatment effect can be estimated by calculating the mean CATE over all observed patients.

A.1.4 Estimation in a Marginalized DAG

For any application of PROTECT, the number of possible cause and proxy variables of patient fitness may be large. Moreover, different research groups investigating the same application may come up with different sets of cause and proxy variables of fitness. A natural question is whether the estimated treatment effect depends on the chosen set of proxies. Under the assumption that there is a finite number of potential cause and proxy variables of fitness, we will show that omitting some these variables will not necessarily bias the treatment effect estimate. For this we need two things to hold:

1. The DAG with fewer variables is indeed the DAG that is obtained after marginalizing out the unused variables from the full DAG
2. Given the DAG with fewer variables, the conditional average treatment effect is identified from the observed data

We defer the discussion of the second requirement to subsection A.1.4. To demonstrate the first requirement we will consider a hypothetical application of PROTECT to a case where there exist three binary proxies of fitness (w_1, w_2, w_3) and no cause variables for fitness. A research groups tries to estimate the treatment effect but only has measurements of two of the three proxy variables. See Figure A.1.4 for the DAG for this situation.

As the researchers only have samples from

$$p(y, t_x, w_1, w_2) = \sum_{\mathrm{w}=0}^{1} p(y, t_x, w_1, w_2, w_3 = \mathrm{w})$$

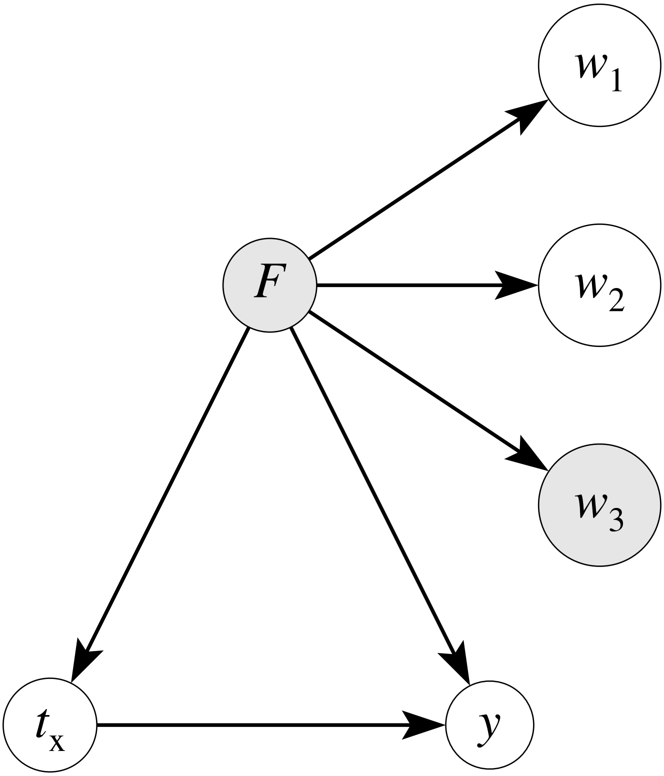


Figure 2: Example DAG where one of the three proxies is not observed.

the target estimand for this study is CATE(w_1, w_2) instead of CATE(w_1, w_2, w_3). If the group had access to the joint distribution $p(y, t_x, w_1, w_2, w_3)$, they could calculate CATE(w_1, w_2) by first summing out w_3 from the joint distribution and then calculating the CATE. The question is whether this will lead to the same estimand as when estimating CATE(w_1, w_2) from samples from $p(y, t_x, w_1, w_2)$ directly. We now show that this is indeed the case.

$$
\begin{aligned}
&\sum_{\mathrm{w}=0}^{1} p(y, t_x, w_1, w_2, w_3 = \mathrm{w}) \\
&\qquad = \sum_{\mathrm{w}=0}^{1} \int_F p(y|t_x, F)p(t_x|F)p(w_1|F)p(w_2|F)p(w_3 = \mathrm{w}|F)p(F)dF \\
&\qquad = \int_F p(y|t_x, F)p(t_x|F)p(w_1|F)p(w_2|F) \sum_{\mathrm{w}=0}^{1} \left[p(w_3 = \mathrm{w}|F)\right] p(F)dF \\
&\qquad =^1 \int_F p(y|t_x, F)p(t_x|F)p(w_1|F)p(w_2|F)p(F)dF \\
&\qquad = p(y, t_x, w_1, w_2)
\end{aligned}
$$

In ^{1} the back-door requirement is still fulfilled so the treatment effect can be estimated. The average treatment effect is calculated by calculating the expectation of the CATE over the proxy variables. As the CATE(w_1, w_2) can be estimated from samples of $p(y, t_x, w_1, w_2)$ and the order of taking the expectation over proxy variables does not matter, the inferred average treatment effect would be the same as first estimating CATE(w_1, w_2, w_3) and then taking the expectation over all three proxy variables. A similar argument can be made for cause variables of F.

A.1.5 Treatment Effect Estimation

As mentioned in the main text, if the joint distribution of observed variables and the latent confounder can be estimated from the observed data, the treatment effect can be estimated. This is because the back-door adjustment formula can be applied using the estimated distributions of survival given treatment and fitness, and fitness given the proxy variables and cause variables of fitness [16, 12], see also equation 4. The crucial question is whether the joint distribution of observed variables and the latent confounder can be correctly estimated from the observed data. Without constraints on the joint distribution, this is not possible. Fortunately, the DAG provides conditional independency constraints on the joint distribution. For instance, the treatment choice is independent of performance score, once the value of fitness is known. However, this is not enough to identify the joint distribution in general and additional assumptions are needed. These assumptions can be provided by the form of error distributions of the observed variables and latent confounder or by functional forms of relationships between these variables.

In PROTECT, parametric forms for the structural equations in the DAG are specified. By assuming parametric models and thus reducing the family of structural causal models under consideration, the question of identification of the treatment effect reduces to whether the parameters for the structural equations can be uniquely estimated from observed data. This is not an easy question to answer in general. If all observed variables and the latent factor followed linear models with Gaussian error distributions, a well-known sufficient condition for uniqueness of the parameters is when there are at least three independent proxies per latent factor [2]. This textbook result of parameter identification is based on comparing the number of unknown parameters in the statistical model with the number of unique entries in the covariance matrix of the observed variables. The latent factor fitness in our DAG has four dependent variables (two proxies, the treatment and overall survival). Treatment and survival also have a direct dependency relationship that would require one additional parameter to be estimated in a linear Gaussian setting. The requirement of estimating this single extra parameter is offset by having one more observed variable, so this model would still be identified. However, many settings will be more complex than standard linear structural equation models, and analytical identification proofs are often intractable to obtain [20]. Instead, empirical checks can be performed to see if there is any evidence of non-uniqueness of the treatment effect estimate given the observed data and modeling assumptions. In a Bayesian parameter estimation setting, this amounts to checking the posterior distribution over parameters for multi-modality.

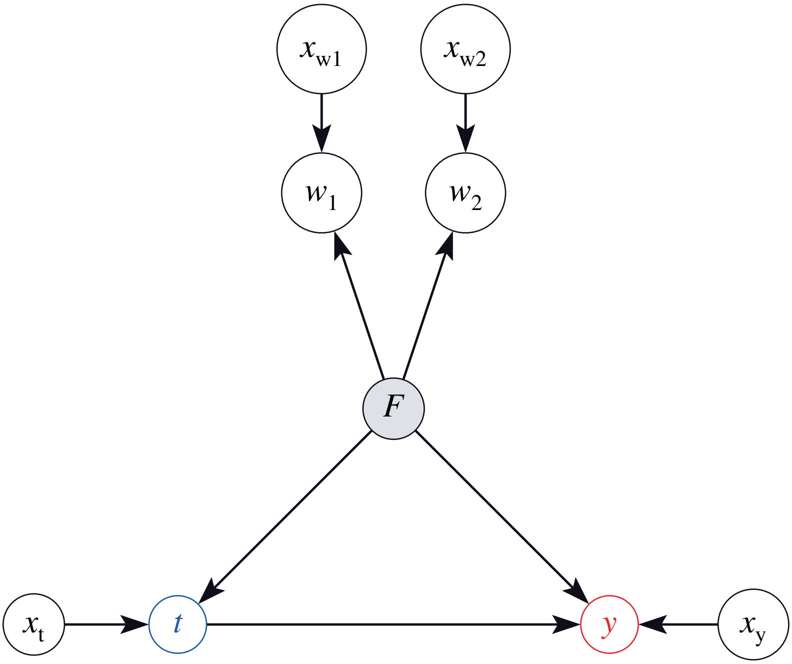


Figure 3: Exeample of a DAG where PROTECT could be applied

A.1.6 Model Selection

In PROTECT, the joint distribution of observed variables and unobserved variables is estimated by specifying parametric distributions for all structural equations implied by the DAG. Different parameterizations may lead to different inferences regarding the treatment effect. To reduce the dependence of the treatment effect estimate on these choices, we now introduce a data-driven model selection procedure. We describe it for a general DAG with a latent confounder with proxy variables and causes.

In figure 3 we present a general latent confounder model with two proxies w_1, w_2, a treatment variable t and outcome y. Each observed variable has a possibly empty set of control variables $\mathbf{x}_{(.)}$ and the sets of control variables are allowed to overlap. Causes of the latent factor F are permitted but are irrelevant to the model selection procedure so they are omitted here. The latent factor F is a cause of the proxies, treatment and outcome, and is the only confounder of t and y. The proxies, treatment and outcome are assumed to be random variables conditional on their parents in the graph. This graph is necessarily an abstraction of complicated clinical and biological processes. In reality, the process that is responsible for treatment decisions and outcomes may be better described with a multi-dimensional latent factor F. Considering this multi-dimensional F, it is likely that there are dimensions of F that are related to treatment but not to the outcome (F_t) and vice-versa (F_y). The dimensions of F that are related to both treatment and outcome are denoted $F_{t,y}$ and constitute the true confounder. See Figure 4. Possible clinical interpretations of these dimensions are:

- F_t: information that the pre-treatment variables provide that is thought to be relevant to efficacy of treatment (and is used as such to select patients for treatment), but in reality holds no information on treatment efficacy or outcome

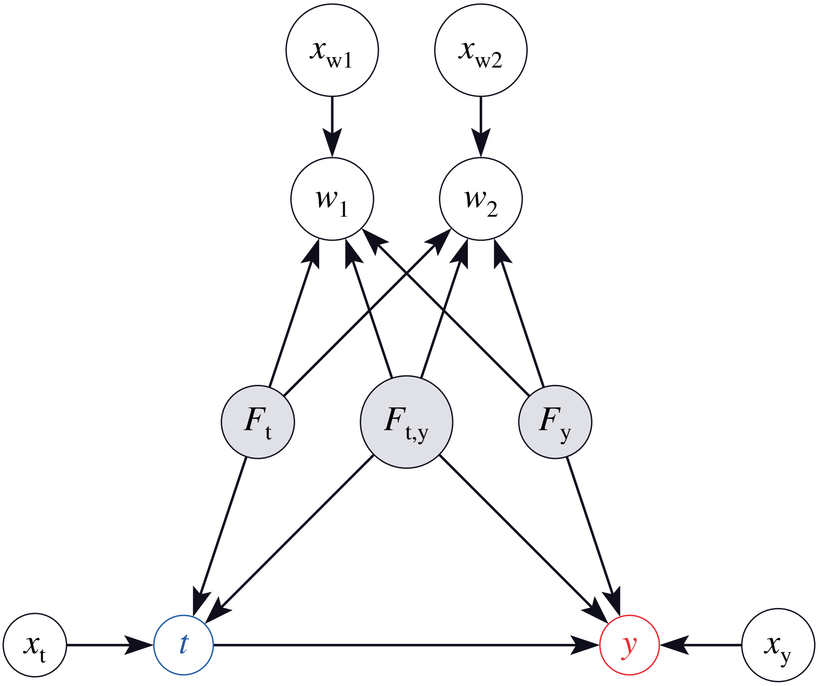


Figure 4: A more complex true data generating process may underlie the observed data which is modeled as in Figure 3 F_t is a latent factor that only influences the proxy variables and the treatment assignment. F_y is a latent factor that only influences the proxy variables and the outcome. $F_{t,y}$ is a latent factor that influences the proxy variables and both the treatment and the outcome. $F_{t,y}$ is the only confounder.

- F_y: information that the pre-treatment variables have on the outcome that is not known to the physician (this could be treatment effect modification that is not known to the physician), or otherwise is not used to make treatment decisions.
- $F_{t,y}$: information that the pre-treatment variables have on the outcome or treatment efficacy and that is utilized in the treatment decision process

To estimate the treatment effect, the modeled latent factor $\hat{F}$ should contain as much information on $F_{t,y}$ as possible. Both F_t and F_y are irrelevant for the estimation of the treatment effect in terms of bias.

When the parameterization of the model has limited expressiveness (e.g. $\hat{F}$ is a one-dimensional latent variable), it is not guaranteed that the model parameters that maximize the joint likelihood of the observed data will learn an $\hat{F}$ that has information about $F_{t,y}$. For example, when the mutual information between the pre-treatment variables and treatment is much higher than the mutual information between pre-treatment variables and outcome, it could be that a model with a single dimensional $\hat{F}$ will learn $\hat{F} = F_t$. Indeed, this is not an unrealistic scenario. The treatment selection process is mostly determined by the pre-treatment variables. Outcomes like overall survival have much higher intrinsic randomness given these pre-treatment variables as they depend on 1. the true biological state of the patient and the tumor that may never be fully known and 2. random events in the future that have not occurred at the time of the treatment choice. In the following paragraph we introduce a set of model checks based on observed data to evaluate whether a given parameterization of the graphical model in the DAG is compatible with

Hyperparameter Selection When there are multiple choices for parameterizing the joint model summarized in hyperparameter ψ (e.g. number of dimensions for $\hat{F}$, parameterization choices for any of the conditional distributions, including priors for parameters), each value of ψ will imply a posterior distribution $p_{\text{post},\psi}(\theta) = p_\psi(\theta|D_{\text{train}})$ over global parameters θ after conditioning on training data D_{train}. To infer if a chosen hyperparameter ψ is compatible with the assumptions on the data generating mechanism in the DAG, a number necessary constraints implied by the DAG can be checked on held-out data D_{test}. Let $p_\psi(y^{(i)}|t^{(i)}, \mathbf{w}^{(i)}, \mathbf{x}^{(i)}, \theta)$ be the predictive density for the outcome of observation i, conditional on treatment, proxies and control variables for the outcome, of the model with hyperparameter ψ and global parameter value θ. Let $l_\psi(y^{(i)}|t^{(i)}, \mathbf{w}^{(i)}, \mathbf{x}^{(i)})$ be the log point-wise predictive density for observation i from the test data, given hyperparameter value ψ: $l_\psi(y^{(i)}|t^{(i)}, \mathbf{w}^{(i)}, \mathbf{x}^{(i)}) = \log \mathbb{E}_{\theta \sim p_{\text{post},\psi}(\theta)} p_\psi(y^{(i)}|t^{(i)}, \mathbf{w}^{(i)}, \mathbf{x}^{(i)}, \theta)$. Here the expectation over θ can be approximated e.g. by using monte-carlo samples of θ from the posterior distribution. Let $L_\psi(y^{(i)}|t^{(i)}, \mathbf{w}^{(i)}, \mathbf{x}^{(i)})$ be the expectation of the log point-wise predictive likelihood (*elpd*) [22] over the held-out data, and let the *elpd* for different conditional distributions be similarly defined, then for t, y and proxies w_j the following minimal criteria should hold:

$$L_\psi(w_j|y, t, \mathbf{w}_{-j}, \mathbf{x}_{w_j}) = \mathbb{E}_{F \sim p_\psi(F|y,t,\mathbf{w}_{-j},\mathbf{x}_{w_j})} p_\psi(w_j|\mathbf{x}_{w_j}, F) > L(w_j|\mathbf{x}_{w_j}) \tag{5}$$

$$L_\psi(t|y, \mathbf{w}, \mathbf{x}_t) = \mathbb{E}_{F \sim p_\psi(F|y,\mathbf{w},\mathbf{x}_t)} p_\psi(t|\mathbf{x}_t, F) > L(t|\mathbf{x}_t) \tag{6}$$

$$L_\psi(y|t, \mathbf{w}, \mathbf{x}_y) = \mathbb{E}_{F \sim p_\psi(F|t,\mathbf{w},\mathbf{x}_y)} p_\psi(y|t, \mathbf{x}_y, F) > L(y|t, \mathbf{x}_y) \tag{7}$$

In words this means that for each observed variable, the predictive likelihood of the model with the latent variable that conditions on the other observed variables should be greater than a corresponding baseline model that conditions only on the direct parents in the DAG, excluding the latent factor. To calculate the *elpd* for the baseline models, a separate regression model was formulated for each target using the same linear formula, outcome distribution and priors as in the DAG model, but without the latent factor. If the predictive likelihood of the model with latent factor is higher than the baseline regression model without the latent factor, then the latent factor effectively transmits information between observed variables in the model indexed by ψ. However, this still does not guarantee that the inferred $\hat{F}$ is the actual confounder $F_{t,y}$. If we select a hyperparameter $\psi = \psi_t$ that leads to $\hat{F} = F_t$ then equations 5 may still be satisfied. We can formulate more strict requirements. If we mutilate the DAG in Figure 4 by removing $F_{t,y}$ and F_y and their outgoing arrows, the following equality can be shown by applying the conditional independence of $F_t|t$ of $y|t$ implied by the mutilated DAG:

$$\begin{aligned} p_{\psi_t}(w_j|y, t, \mathbf{w}_{-j}, \mathbf{x}_{w_j}) &= \mathbb{E}_{F_t \sim p_\psi(F_t|y,t,\mathbf{w}_{-j},\mathbf{x}_{w_j})} p_\psi(w_j|\mathbf{x}_{w_j}, F_t) \\ &= \mathbb{E}_{F_t \sim p_\psi(F_t|t,\mathbf{w}_{-j},\mathbf{x}_{w_j})} p_\psi(w_j|\mathbf{x}_{w_j}, F_t) \\ &= p_{\psi_t}(w_j|t, \mathbf{w}_{-j}, \mathbf{x}_{w_j}) \end{aligned}$$

Equivalently, if we select a hyperparameter $\psi = \psi_y$ that leads to $\hat{F} = F_y$ then $L_{\psi_y}(w_j|y, t, \mathbf{w}_{-j}, \mathbf{x}_{w_j}) = L_{\psi_y}(w_j|y, \mathbf{w}_{-j}, \mathbf{x}_{w_j})$. Note that this equality does not hold if either $F_{t,y}$, or the combination of F_t and F_y are in the DAG. We can now define a necessary condition for the $\hat{F}$, estimated from the observed data, not to be independent of $F_{t,y}$. For all w_j:

$$L_{\psi_{t,y}}(w_j|y, t, \mathbf{w}_{-j}, \mathbf{x}_{w_j}) > L_{\psi_{t,y}}(w_j|t, \mathbf{w}_{-j}, \mathbf{x}_{w_j}) \tag{8}$$

$$L_{\psi_{t,y}}(w_j|y, t, \mathbf{w}_{-j}, \mathbf{x}_{w_j}) > L_{\psi_{t,y}}(w_j|y, \mathbf{w}_{-j}, \mathbf{x}_{w_j}) \tag{9}$$

Hyperparameter settings for which one of these conditions does not hold for one of the proxies should be rejected. Note that this does not rule out the possibility that $\hat{F}$ is some function of F_t and F_y and is still not related to $F_{t,y}$.

A.2 Methods for stage III NSCLC application

In this subsection more implementation details for the application of PROTECT to stage III Non-Small Cell Lung cancer (NSCLC) are provided.

A.2.1 Missing Data

Proxies of fitness Values of proxies of fitness are assumed to be missing at random conditional on the observed variables. We further assumed independent priors for the proxy missingness mechanism. Together with the missing at random assumption, this makes the missingness model ignorable, meaning that it does not contribute to the estimation of the other parameters. We therefore did not model missingness in proxies. Since we model the joint distribution of proxies, treatment and survival, marginalization of the missing values of proxy variables is trivial by not adding terms to the likelihood for values that are not observed.

Weight loss Roughly 15% of the values for weight loss were missing. Though weight loss is a known prognostic factor and is a standard part of the pre-treatment history taking, the physician may sometimes forget to ask about it, or forget to note it down in the electronic health record. It is likely that high values of weight loss have a higher probability of being registered. A patient with high weight loss may self report it and the physician is more likely to report it, as it is more notable. This renders the missingness mechanism for weight loss "Missing Not at Random". When the missingness in a covariate is dependent on the value of the covariate (but not on the outcome), estimating parameters of a regression model for the outcome using the complete cases is not biased by the missingness, while imputation may lead to bias in the estimation of the treatment effect [23]. A problem with the complete case method here is that allthough the conditional treatment effect may be validly estimated, the average treatment effect may be different in the population of interest than in the complete case population, as the populations may differ with respect to the distribution of weight loss. If the effect of treatment depends on weight loss (according to our ADMG, either through the effect of tumor behavior on survival, or through the effect of tumor behavior on fitness), the average treatment effect estimated from the complete cases with respect to weight loss is a biased estimator for the average treatment effect in the target population. Through a sensitivity analysis presented in section B.2.2 we assess what the effect of this missingness on the estimated average treatment effect may be.

A.2.2 Pre-processing, Parametric Models and Priors

In this subsection we describe details on the data pre-processing, the parametric forms of each conditional distribution and the priors for all parameters. The Gaussian distribution parameterized with mean μ and standard deviation σ is denoted $\mathcal{N}(\mu, \sigma)$, the Half-Normal distribution with standard deviation σ is denoted $\text{HN}(\sigma)$, the Bernoulli distribution with probability p is denoted as $\text{Bern}(p)$, the Uniform distribution over values between a and b is denoted $\mathcal{U}(a, b)$

Covariate pre-processing In order to be able to use the same distribution to model each of the proxies, proxy variables were binarized. Dichotomization of variables is not generally recommended due to the loss of information. However, as our primary goal is to identify the conditional average treatment effect, we expect the potential loss of information from the dichotomization to be small compared to the extra modeling challenges when dealing with proxies with different outcome distributions. Our exact variable definitions were:

- ECOG performance score: 1 or higher vs 0 (higher means worse overall health according to the ECOG standard [15])
- eGFR: less (or greater) than 60 milliliter / minute / $1.73m^2$

The first value (1 for ECOG, and less than 60 for eGFR) was dummy coded as 0, and the second value was dummy coded as 1. There were only two patients with stage IIIC, therefore they were grouped with IIIB patients. Histology types were grouped as Adenocarcinoma, Squamous Cell Carcinoma or other, missing values were classified as other. Weight loss was defined as any weight loss of $\geq 3\%$ of the original weight in the 6 months preceding the start of follow-up. Age was scaled to zero mean and standard deviation 1. Details on the definition of the treatment variable and outcome are given in the main text.

Survival The Cox-proportional hazards model is characterized by a partial likelihood that leaves the baseline hazard unspecified. Performing Bayesian inference with a proportional hazards model will require specifying a likelihood for the baseline hazard. We chose a parametric survival model, using the Adaptive Power Generalized Weibull (APGW) distribution as described in [4]. The APGW has two parameters that control the shape of the baseline hazard function. With these two parameters, the APGW can model a wide range of baseline hazard function shapes. The APGW accommodates non-proportional hazards effects by letting one or more of the shape parameters depend on covariates. It can be parameterized as an accelerated failure-time model or as a (proportional) hazards model. Using the proportional hazards formulation of the APGW makes it very similar to Cox-proportional hazards regression, but with a parametric baseline hazard function. The β coefficients in this version have the same interpretation as the parameters in the familiar Cox-proportional hazards regression, namely log hazard ratios. The reference value for the treatment effect is presented as a hazard ratio [1]. This value assumes a proportional hazard model for overall survival. Therefore, we chose to parameterize the outcome model similarly using a linear model for the log-hazard ratio in the proportional hazards formulation of the APGW. Potential treatment effect modification for variables was accommodated by adding parameters for product terms of the variable and the treatment. The APGW was parameterized as follows:

$$\begin{aligned}
\beta(t_x, F, X_y) &= \beta_0 + F\beta_{F\to y} + X_y\beta_{X_y\to y} + && (10)\\
&\quad t_x(\beta_{t_x\to y} + F\beta_{F*t_x\to y} + X_y\beta_{X_y*t_x\to y}) && (11)\\
\{\text{time}, \text{deceased}\} &\sim \text{APGW}(0, \beta(t_x, F, X_y), \alpha_0, \nu_0) && (12)\\
\beta_{X_y\to y}, \beta_{t_x\to y} &\sim \mathcal{N}(0, 2.5) && (13)\\
\beta_0, \alpha_0 &\sim \mathcal{N}(0, 5) && (14)\\
\nu_0 &\sim \mathcal{U}(-5, 5) && (15)\\
\beta_{F\to y} &\sim \mathcal{N}(0, \sigma_{F\to y}) && (16)\\
\beta_{X_y*t_x\to y}, \beta_{F*t_x\to y} &\sim \mathcal{N}(0, 0.1) && (17)
\end{aligned}$$

Here, $t_x = 1$ is concurrent chemoradiation, F is the inferred latent fitness, X_y are the other direct parents of survival in the ADMG, $\sigma_{F\to y}$ is a hyperparameter that was determined per the model selection method in subsection A.1.6. Using this parameterization, we can now express the CATE (equation 1) for patient i in terms of a log hazard ratio β:

$$\begin{aligned}
\text{CATE}^{(i)} &= \beta(t_x = 1, F^{(i)}, X_y^{(i)}) - \beta(t_x = 0, F^{(i)}, X_y^{(i)}) && (18)\\
&= \beta_{t_x\to y} + F^{(i)}\beta_{F*t_x\to y} + X_y^{(i)}\beta_{X_y*t_x\to y} && (19)
\end{aligned}$$

The average treatment effect (ATE) is estimated with the mean CATE over the observed population.

Latent Factor The latent factor F is modeled using a conditional Gaussian distribution with fixed standard deviation 1, followed by the logistic function (also known as the sigmoid function, denoted σ). The logistic function was used to fix the overall scale of the latent factor to prevent compensatory scaling of F when adjusting hyperparameters that control the magnitude of the effect of F on t_x or y.

$$
\begin{aligned}
\mu_F &= X_F \beta_{X_F \to F} \\
F &\sim \sigma(\mathcal{N}(\mu_F, 1)) \\
\beta_{X_F \to F} &\sim \mathcal{N}(0, 2.5)
\end{aligned}
$$

In the conditional mean function for F, the control variables X_F were pre-centered to have zero mean.

Proxies We parameterize the proxies as conditionally independent Bernoulli distributed random variables. The conditional probability was implemented using the logistic link function. For each proxy w_j:

$$
\begin{aligned}
\eta_{w_j} &= F\beta_{F \to w_j} - \mu_{wj} \\
w_j &\sim \text{Bern}(\sigma(\eta_{wj})) \\
\beta_{F \to w_j} &\sim \text{HN}(2.5) \\
\mu_{w_j} &\sim \mathcal{N}(0, 2.5)
\end{aligned}
$$

The prior $HN(2.5)$ implements the assumption that a higher value of latent fitness leads on average to better values for the proxy variables of fitness.

Treatment The treatment indicator was modeled as a Bernoulli distributed random variable, with a linear model and the logistic link function, similar to the proxy variables.

$$
\begin{aligned}
\eta_{t_x} &= F\beta_{F \to t_x} + X_{t_x}\beta_{X_{t_x} \to t_x} - \mu_{t_x} \\
t_x &\sim \text{Bern}(\sigma(\eta_{t_x})) \\
\beta_{F \to t_x} &\sim \text{HN}(\sigma_{F \to t_x}) \\
\mu_{t_x} &\sim \mathcal{N}(0, 2.5)
\end{aligned}
$$

X_{t_x} are all observed variables that are direct parents of treatment in the ADMG, except for F. $\sigma_{F \to t_x}$ is a hyperparameter that was determined per the model selection method in subsection A.1.6.

Structural Causal Model The entire generative model can now be formalized in a single structural causal model. Let $X^{(.)}$ denote a $Nd^{(.)}$ design matrix for N patients with $d^{(.)}$ control variables for different control variable sets, $\beta_{(.)}$ global parameters, APGW(.) the 4-parameter Adapted Power Generalized Weibull distribution [4], w_j binary proxies, t_x treatment variable, y the survival outcome consisting of positive real number indicating time, and a binary indicator for deceased or censored.

$$
\begin{aligned}
\mu_F &= X^F \beta_{X^F \to F} \\
F &\sim \sigma(\text{N}(\mu_F, 1)) \\
\eta_{w_j} &= F\beta_{F \to w_j} - \mu_{w_j} \\
w_j &\sim \text{Bern}(\sigma(\eta_{w_j})) \\
\eta_{t_x} &= F\beta_{F \to t_x} + X_{t_x}\beta_{X_{t_x} \to t_x} - \mu_{t_x} \\
t_x &\sim \text{Bern}(\sigma(\eta_{t_x})) \\
\beta_y &= F\beta_{F \to y} + X_y\beta_{X_y \to y} + t_x(\beta_{t_x \to y} + F\beta_{F * t_x \to y} + X\beta_{X * t_x \to y}) + \beta_0 \\
y &\sim \text{APGW}(0, \beta_y, \alpha_0, \nu_0)
\end{aligned}
$$

Effect Modifiers Predicting treatment effects for new patients requires the estimation of the conditional average treatment effect (CATE). In the case of a binary treatment variable t_x and covariates x, w, the CATE is defined as:

$$\text{CATE}(x, w) = \mathbb{E}\left[y|\text{do}(t_x = 1), x = x, w = w)\right] - \mathbb{E}\left[y|\text{do}(t_x = 0), x = x, w = w\right] \tag{20}$$

To evaluate whether some variables are effect modifiers (i.e. the treatment effect differs between different values of this variable), we employ the definition of a conditional effect modifier by VanderWeele [21]. For covariates x, w and treatment t_x, w is said to be an effect modifier of t_x conditional on x if for some value of x there exist two values $w_1 \neq w_2$ of w for which $\text{CATE}(x, w_1) \neq \text{CATE}(x, w_2)$. Note that the estimation of a treatment effect modifier is dependent on the scale on which the CATE is measured. For a linear log-hazard ratio model (e.g. the Cox proportional hazards model or the proportional hazards APGW), this reduces to the familiar statistical interaction term β_{w*t_x}:

$$\beta(t_x, x, w) = \beta_{t_x} t_x + \beta_x x + \beta_w w + \beta_{w*t_x} w t_x \tag{21}$$

Filling in equation 21 in 20 will lead to the familiar result that the conditional average treatment effect is the average treatment effect plus $w\beta_{w*t_x}$. This treatment interaction term quantifies the difference in treatment effects between $w = 0$ and $w = 1$. Due to the linearity of this model in 21, the effect modification of w is the same for all values of x. In our log-hazard model formulation, effect modification is similarly modeled by adding product terms between treatment and control variables, but also between treatment and the latent factor F. Since the conditional distribution of $p(F|X, W)$ is non-linear in X and W, effect modification by proxies W and control variables X can no longer be equated to the interaction terms in the linear log-hazard model, and the effect modification of some variable x_i may depend on the value of other control variables and proxy variables.

We can estimate potential effect modification by plugging our log-hazard model $\beta(t_x, F, X_y)$ defined in 18 in equation 20:

$$\begin{aligned} \text{CATE}(X, W) &= \underset{F \sim p(F|X,W)}{\mathbb{E}}[\beta(t_x = 1, X, F)] - \underset{F \sim p(F|X,W)}{\mathbb{E}}[\beta(t_x = 0, X, F)] \qquad (22) \\ &= \underset{F \sim p(F|X,W)}{\mathbb{E}}[\beta(t_x = 1, X, F) - \beta(t_x = 0, X, F)] \qquad (23) \end{aligned}$$

This will estimate the log-hazard ratio between concurrent and sequential treatment as a function of X and W. We have omitted the decency of these quantities on the global parameters θ here. As said before, the effect modification of a variable can depend on the value of the other variables due to the non-linearities in our model. To summarize the average effect modification for a unit change in a variable we use partial dependence functions [8]. Specifically we look at the average change in CATE for a unit difference in a variable x_j, averaged over the marginal distribution of other variables $\mathbf{x}_{-j}$.

$$\text{EM}(x_j) := \mathbb{E}_{\mathbf{x}_{-j} \sim p(\mathbf{x}_{-j})}\left[\text{CATE}(x_j = 1, \mathbf{x}_{-j}) - \text{CATE}(x_j = 0, \mathbf{x}_{-j})\right] \tag{24}$$

Where the expectation is taken over the observed data.

A.2.3 Model selection

As the treatment variable and the outcome variable follow different distributions there is no way to express the parameters of these models on a single scale. Specifically, the log odds ratio of fitness to treatment ($\beta_{F \to t_x}$) is incommensurable with the log hazard ratio of fitness to survival ($\beta_{F \to y}$). This creates a problem with specifying the right scale for the prior distributions of these two parameters. If the scale of the priors for one of the parameters is greater than the other, parameters that maximize the joint likelihoood can be biased towards modeling the variable with the prior with the greater scale. Therefore, the prior standard deviations on the parameters from F to y and from F to t_x were treated as hyperparameters. Values for the prior standard deviations were $\{0.1, 1.0, 2.5, 10.0, 100.0\}$, resulting in a grid of 125 hyperparameter combinations. We used 5-fold cross validation to select hyperparameters that were not refuted by the assumptions in the DAG as outlined in A.1.6. In a final inference step we follow a Bayesian Model Averaging approach by formulating a mixture model over all acceptable hyperparameter settings ψ_j.

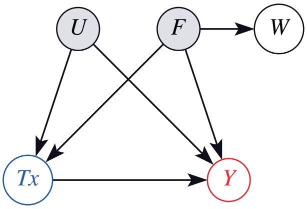
$$p_{\text{BMA}}(\theta) = \sum_j p_{\psi_j}(\theta|D_{\text{full}}) p(\psi_j|D_{\text{full}}) \tag{25}$$

Where each ψ_j corresponds to an acceptable setting of the hyperparameters. This mixture model was evaluated once on the full dataset where global parameters and model weights were inferred jointly, resulting in our final model estimate.

A.2.4 Implementation

All probabilistic models were implemented using the open source probabilistic programming language NumPyro [18], version 0.4.1 with JAX [3] version 0.2.7 as a back-end. We used the No-U-Turn Hamiltonian Monte Carlo sampling algorithm to simulate 7500 samples per chain (following 2500 warm-up samples) in 4 independent chains from the posterior distribution for each hyperparameter setting. For posterior predictive densities on held-out data we numerically integrated the densities with respect to F by using a fixed grid of points for $\hat{F}_\epsilon$, the noise term of $\hat{F}$. Between -5.0 and 5.0, 250 equally spaced values for $\hat{F}_\epsilon$ were used. From the original samples over global parameters, 1500 samples equally divided over the chains were used. The samples of global parameters were selected by slicing the original chain to reduce auto-correlation between the samples. After applying the model selection procedure we estimated the final model using 12 independent chains with 7500 samples per chain (following 2500 warm-up samples). Model evaluation was performed in R, version 4.0.3. The code that implements the statistical models will be made freely accessible online. Due to privacy regulations the clinical data cannot be made available.

Figure 5: Causal Directed Acyclic Graph for sensitivity analysis of omitted confounder. A simplified DAG is used to present the sensitivity analysis.



B Sensitivity Analyses

B.1 Omitted Confounder

B.1.1 Methods

We performed a sensitivity analysis to asses the robustness of our inferences with respect to a potential unobserved confounder. Specifically, we test the effect of an unobserved confounder U on the point estimate of the average treatment effect (ATE). The latent factor U is assumed to be a direct cause of both treatment and survival, independent of the latent factor fitness, see Figure B.1.1.

Assuming a true data generating mechanism (under the sensitivity parameters $\gamma_{U\to y}, \gamma_{U\to t_x}$):

$$\begin{aligned}\beta_y = \beta_0 &+ F\beta_{F\to y} + X_y\beta_{X_y\to y}\\ &+ t_x(\beta_{t_x\to y} + F\beta_{F*t_x\to y} + X_y\beta_{X*t_x\to y})\\ &+ \gamma_{U\to y}U\end{aligned}$$

Where β_y is the log-hazard ratio. Whereas we estimated the treatment effect using the equivalent model for survival but omitting the term from U to y.

As in this hypothetical setting of the sensitivity analysis we did not condition on all confounders when $|\gamma_{U\to t_x}| > 0$ and $|\gamma_{U\to y}| > 0$, the estimated treatment effect is a biased estimate of the true treatment effect. For different settings of sensitivity parameters $\gamma = \{\gamma_{U\to tx}, \gamma_{U\to y}\}$ we re-estimated the posterior distribution over global parameters, this time including U as an additional latent factor with fixed coefficients to treatment and survival, as specified by γ. In order to be able to clinically reason about the likelihood of the existence of an omitted confounder U that would be strong enough to alter the conclusions of our modeling, we need to be able to compare the effects of this hypothetical unobserved confounder on treatment ($\gamma_{U \rightarrow t_x}$) and survival ($\gamma_{U \rightarrow y}$), relative to the effects of the modeled confounder for fitness F ($\beta_{F \rightarrow t_x}, \beta_{F \rightarrow y}$). To make sure that the comparison of parameters was valid we fixed the standard deviation of U to 1, and we re-scaled the parameters $\beta_{F \rightarrow t_x}, \beta_{F \rightarrow y}$ to the values they would have if F were also scaled to have a standard deviation of 1. We investigate the effect of potential unobserved confounding on the treatment effect estimate for all combinations of $\gamma_{U \rightarrow y} \in \{-1, -0.5, 0, 0.5, 1\}\beta_{F \rightarrow y}$ and $\gamma_{U \rightarrow t_x} \in \{0, 0.5, 1\}\beta_{F \rightarrow t_x}$.

B.1.2 Results

The results of the sensitivity analysis are presented in Figure 6. If there were a confounder that was more than half as strong as the modeled latent confounder $\hat{F}$ but opposite in sign of survival, the point estimate of the ATE would be greater than the estimate from the RCTs [1]. The interpretation of this latent confounder is that it increases the likelihood of being treated, but reduces the likelihood of survival. A clinical example may be that patients who have more aggressive tumors are treated more aggressively in order to improve survival outcomes.

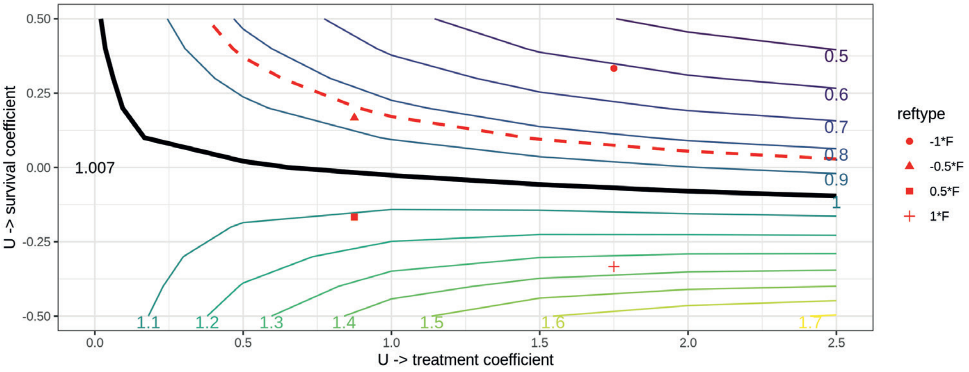


Figure 6: Results of sensitivity analysis. For different combinations of sensitivity parameters $\gamma_{U \rightarrow y}$ (y-axis) and $\gamma_{U \rightarrow t_x}$ (x-axis) we re-estimated the model, with the additional latent variable U. Results for the estimated average treatment effect are indicated by the level lines, marked by the point estimate of the hazard ratio for overall survival of concurrent versus sequential chemoradiation. The black line indicates the null-effect (hazard ratio $=1$). The red dotted line indicates the treatment effect reported in the meta-analysis of randomized controlled trials (hazard ratio $=0.84$) [1]. Four reference points are added to gauge how the sensitivity parameters relate to the parameters of the estimated latent confounder F to treatment and survival. The reference point $1 * F$ corresponds to the value of $\hat{\beta}_{F \rightarrow y}$ and $\hat{\beta}_{F \rightarrow t_x}$ in the original model. $0.5 * F$ is the point where these coefficients are both divided by two. $-1 * F$ is the point where the sign of $\hat{\beta}_{F \rightarrow y}$ is changed, and $-0.5 * F$ is defined analogous to $0.5 * F$

B.2 Missing Weight Loss

We now describe the sensitivity analysis for the effect of different missingness mechansisms for weight loss on the estimate of the average treatment effect.

B.2.1 Methods

Let M denote the missingness indicator for weight loss X, then $p(M = 1|X = x)$ denotes the probability of missingness, conditional on the value of X. The central assumption of this sensitivity analysis is that the missingness in weight loss is random conditional on the value of weight loss itself. Specifically, that the probability of missingness is lower when weight loss is present.

$$p(M = 1|X = 1) \leq p(M = 1|X = 0) \tag{26}$$

Applying this inequality to the marginal probability of missingness $p(M = 1)$ we get:

$$\begin{aligned} p(M = 1) &= p(M = 1|X = 1)p(X = 1) + p(M = 1|X = 0)(1 - p(X = 1)) \\ &\geq p(M = 1|X = 1)p(X = 1) + p(M = 1|X = 1)(1 - p(X = 1)) \\ &= p(M = 1|X = 1) \end{aligned}$$

It follows that $p(M = 1|X = 1) \in [0, p(M = 1)]$. From the observed data we know $p(M = 1)$, the marginal probability of missingness. We introduce sensitivity parameter α to parameterize the range of possible values for $p(M = 1|X = 1)$ compatible with the observed marginal probability of missingness and our assumption 26. We define:

$$p(M = 1|X = 1, \alpha) := (1 - \alpha)p(M = 1) \tag{27}$$

For $\alpha \in [0, 1]$. It then follows that $p(M = 0|X = 1, \alpha) = 1 - (1 - \alpha)p(M = 1)$. In this setup, $\alpha = 0$ represents the boundary case where the missingness in X is completely at random, and $\alpha = 1$ is the boundary case where $p(M = 1|X = 1) = 0$, i.e. X is always observed when $X = 1$.

Let $\pi_{\text{obs}} := p(X = 1|M = 0)$, the probability of weight loss in the observed cases and $\pi^* := p(X = 1)$ denote the unknown true marginal probability of weight loss. Using Bayes rule, we can express π_{obs} in terms of π^*, $p(M = 1)$ and $p(M = 0|X = 1)$:

$$\begin{aligned} \pi_{\text{obs}} = p(X = 1|M = 0) &= \frac{p(X = 1)}{p(M = 0)} p(M = 0|X = 1) \\ &= \frac{\pi^*}{1 - p(M = 1)} p(M = 0|X = 1) \end{aligned}$$

Now we can write:

$$\pi^* = \frac{1 - p(M = 1)}{p(M = 0|X = 1)} \pi_{\text{obs}} \tag{28}$$

Since we cannot estimate $p(M = 0|X = 1)$ from the observed data, we will substitute it with the $p(M = 0|X = 1, \alpha)$ using equation 27, conditional on the sensitivity parameter α.

$$\pi^*_\alpha = \frac{1 - p(M = 1)}{p(M = 0|X = 1, \alpha)} \pi_{\text{obs}} \tag{29}$$

Now we have that given the observed data and the sensitivity parameter α, the unobserved quantity of interest is identified, so we can proceed with the sensitivity analysis.

Re-weighting by weight loss For different values of α we assign weights to patients based on their value for weight loss.

$$w_\alpha(x) := x w^1_\alpha + (1 - x) w^0_\alpha \tag{30}$$

By assigning patients with weight loss a higher weight than patients without weight loss, we can artificially create a population with a greater prevalence of weight loss. Specifically, for each α we define weights such that the weighted sum of the observed values of weight loss is equal to π^*_α.

$$\frac{1}{N}\sum_{i=1}^{N}[w_\alpha(x_i)x_i] = \pi^*_\alpha$$

By further requiring that the sum of weights is equal to the total number of patients,

$$\sum_{i=1}^{N} w_\alpha(x_i) = N$$

the weights are now uniquely defined as:

$$\begin{aligned} w^1_\alpha &= \frac{\pi^*_\alpha}{\pi_{\text{obs}}} \\ w^0_\alpha &= \frac{1 - w^1_\alpha \pi_{\text{obs}}}{1 - \pi_{\text{obs}}} \end{aligned}$$

Using these weights, for each α we calculate a new average treatment effect by taking the weighted mean of the conditional treatment effects:

$$\text{ATE}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} w_\alpha(x_i)\text{CATE}_i \tag{31}$$

B.2.2 Results

The results of this sensitivity analysis are presented in Table B.2.2 and Figure B.2.2. The ATE shows a minor shift in the direction of concurrent being more effective when the dependency of missingness on the actual value of weight loss becomes stronger. This is explained by the fact that patients with weight loss are estimated to have less benefit of concurrent treatment. If missingness were always observed if it was present, this means that the true average value of weight loss is lower in the population than when weight loss is missing completely at random.

α	$p(M=1\|X=1)$	$p(M=1\|X=0)$	π^*_α	w^1	w^0	hazard ratio	CI low	CI high
0.0	0.448	0.170	0.170	1.000	1.000	1.011	0.650	1.534
0.1	0.439	0.153	0.183	1.020	0.983	1.013	0.689	1.630
0.2	0.431	0.136	0.195	1.041	0.967	1.015	0.650	1.536
0.3	0.423	0.119	0.207	1.061	0.950	1.017	0.691	1.636
0.4	0.415	0.102	0.218	1.082	0.934	1.019	0.693	1.639
0.5	0.407	0.085	0.228	1.102	0.917	1.020	0.706	1.672
0.6	0.399	0.068	0.237	1.123	0.900	1.022	0.706	1.673
0.7	0.392	0.051	0.246	1.143	0.884	1.024	0.706	1.675
0.8	0.385	0.034	0.255	1.163	0.867	1.026	0.709	1.681
0.9	0.379	0.017	0.263	1.184	0.850	1.028	0.710	1.682
1.0	0.372	0.000	0.270	1.204	0.834	1.030	0.711	1.684

Results of sensitivity analysis to missing data in weight loss. α: sensitivity parameter, M missingness indicator, X indicator for presence of weight loss, π^*_α true marginal probability of weight loss under sensitivity parameter α, w^1 weight for patients with weight loss, w^0 weight for patients without weight loss, hazard ratio: point estimate of the average treatment effect of concurrent versus sequential chemoradiation for overall survival. CI: 95% credible interval

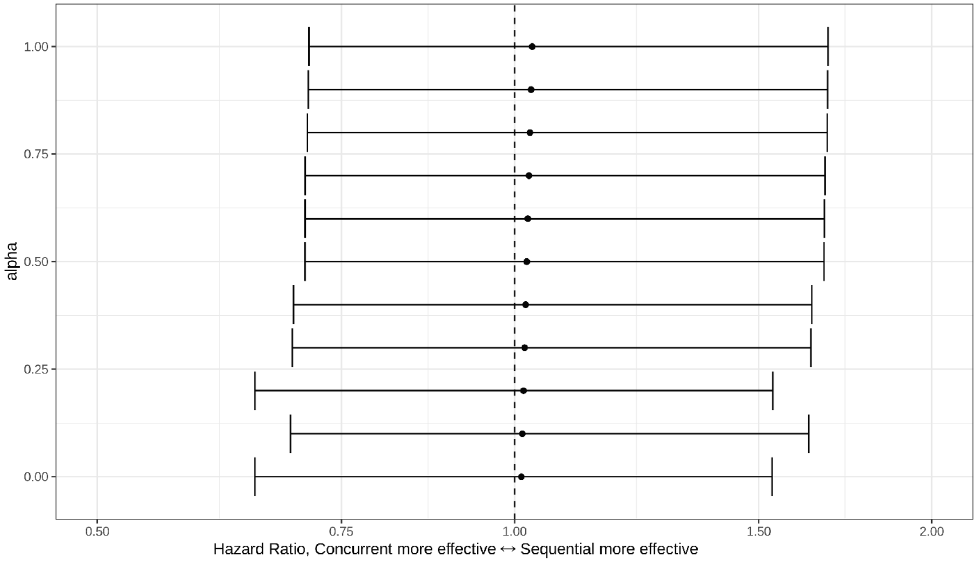


Figure 7: Different average treatment effect values for different values of sensitivity parameter α with 95% credible interval

C Extrapolation to Randomized Trials

C.1 Methods

A central assumption in our study is that the average treatment effect reported in randomized trials is not directly transportable to our population due to differences between the patient populations. We assume however that the conditional treatment effect is transportable. This assumption implies that a patient in our population has the same benefit and harms of treatment as a similar patient in the randomized trials has. Similar here means that they have the same values for the covariates, latent tumor behavior and latent fitness. Using our estimate of the conditional treatment effect we can extrapolate our results to calculate what the estimated average treatment effect would be if our population were more alike the population from the randomized trials.

Matching the RCT Population Average descriptive statistics on the study population of the RCTs (e.g. the mean age) are available from [1]. However, matching our population to theirs is not possible based on these criteria alone. In addition to the published inclusion and exclusion criteria for each RCT included in [1] we should expect hidden confounding between inclusion in the trial and overall survival [7]. Patients included in the randomized trials should be able to undergo all treatment arms. This means that patients who were deemed clinically unfit for concurrent treatment can be expected to be underrepresented in the RCT population. The assessment of this fitness for inclusion in the RCT is essentially the same as the assessment of fitness for concurrent treatment in our real-world population, with the addition of strict inclusion criteria from the trials. See Figure 8 for a DAG that depicts this mechanism. Using the predicted probability of concurrent treatment from the model, we can restrict our population to a subpopulation with higher fitness using different cut-offs for the predicted probability of concurrent treatment.

C.2 Results

We found that the estimated Average Treatment Effect increased in favor of concurrent treatment when restricting the population to higher predicted probability of concurrent treatment, see Figure 9.

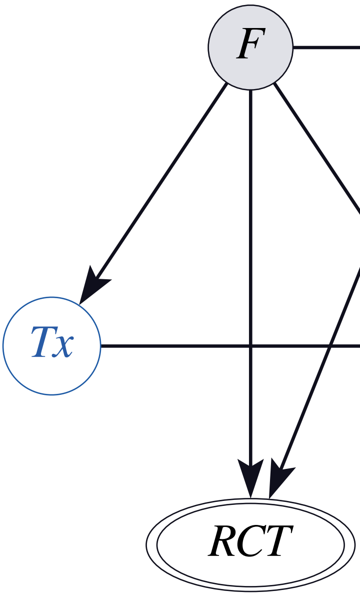


Figure 8: Directed Acyclic Graph of inclusion in RCTs. The same (unobserved) confounding that influences treatment decisions in real-world practice will also influence inclusion in the RCTs. In addition to the unobserved confounder fitness, the concrete exclusion criteria based on proxy measurements (e.g. performance score) also determine inclusion in the RCT

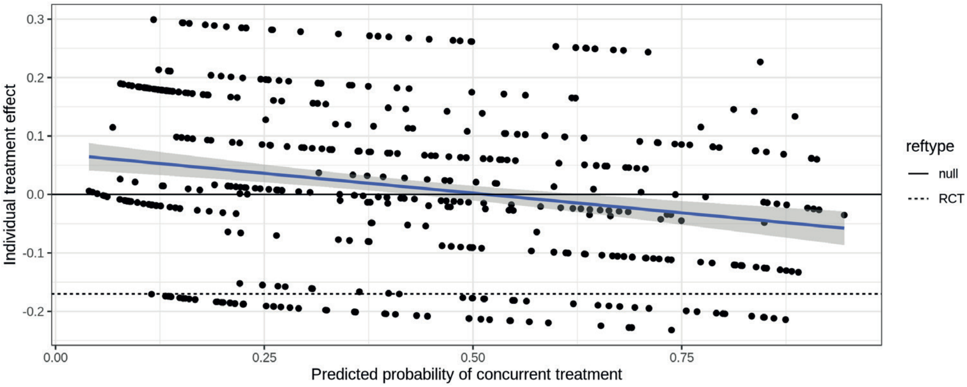


Figure 9: Predicted probability of concurrent treatment versus estimated individual treatment effect per patient. When trying to match the RCT by restricting the population to those with a higher predicted probability of concurrent treatment, the estimated average treatment effect becomes closer to that of the RCTs [1]. Added is a linear regression fit line with 95% prediction confidence band. Reference lines are added for the null-effect (hazard ratio 1) and the effect reported in the randomized trials (hazard ratio 0.84) [1].

	ours	RCT
n	504	1205
age (%)		
<60	143 (28.4)	519 (43.1)
60-64	85 (16.9)	225 (18.7)
65-69	97 (19.2)	270 (22.4)
>=70	179 (35.5)	189 (15.7)
missing	0 (0.0)	2 (0.2)
histology (%)		
adeno	200 (39.7)	395 (32.8)
squamous	194 (38.5)	549 (45.6)
other	89 (17.7)	256 (21.2)
missing	21 (4.2)	5 (0.4)
male sex (%)	300 (59.5)	921 (76.4)
stage (%)		
I	0 (0.0)	6 (0.5)
II	0 (0.0)	12 (1.0)
IIIA	268 (53.2)	441 (36.6)
IIIB	232 (46.0)	735 (61.0)
IIIC	2 (0.4)	0 (0.0)
IV	0 (0.0)	3 (0.2)
missing	2 (0.4)	8 (0.7)

Table 1: Baseline characteristics of participants in our study (ours) and in the meta-analysis of randomized trial (RCTs) [1]

D Results

D.1 Cohort

Patients were recruited from 9 different hospitals in the Utrecht region of the Netherlands. Specifically, the University Medical Center Utrecht, the Antonius Hospital, locations Nieuwegein and Woerden, the Meander Medical Center Amersfoort, the Diakonessen Hospital, Utrecht, the Beatrix Hospital, Gorinchem, the Haaglanden Medical Center, The Hague, the Amsterdam Medical Center, Amsterdam, the Gelderse Vallei Hospital, Ede, and the Antoni van Leeuwenhoek Hospital. These hospitals represent a rich ensemble of university hospitals, specialized oncological hospitals and both large and small secondary care institutions.

All patients were referred to the University Medical Center Utrecht for radiotherapy, but their care was coordinated by the physician in the referring hospital.

Baseline data of our cohort and that of the published meta-analysis of randomized clinical trials [1] are presented in Table 1. Percentages were calculated on the available data. Note that our cohort is older, has worse performance scores and is less frequently treated with concurrent therapy.

D.2 Supplemental Tables

Table 2: Statistics of MCMC samples of parameters of the final bayesian model average. The variables F_mu, F_sd and b_txbinary_y_marginal are not parameters of the model, but are deterministically calculated from the parameters, and are included for reference. The parameters pmodel[0], pmodel[1] and pmodel[2] are the model probabilities for the three selected hyperparameter settings: $\sigma_{F \to t_x} = 2.5$, $\sigma_{F \to y} \in \{0.1, 1.0, 2.5\}$. The parameters are reported on the scale of the linear model (i.e. log odds ratios for the binary proxies and treatment, and log hazard ratios for the outcome survival). Given that a higher hazard means a worse overall survival, the negative sign of b_F_y indicates that higher fitness leads to better overall survival. At the same time the positive sign of b_F_tx indicates that higher fitness leads to a higher chance of concurrent treatment

	mean	sd	hdi_2.5%	hdi_97.5%	mcse_mean	mcse_sd
b_F_eGFRunder60	-2.300	0.878	-4.045	-0.631	0.004	0.003
b_F_ecogbinary1	-2.047	0.575	-3.186	-0.940	0.004	0.003
b_F_tx	6.157	1.486	3.296	9.061	0.013	0.009
b_F_y	-1.171	0.731	-2.511	0.125	0.006	0.005
b_Ftx_y	-0.122	0.354	-0.814	0.572	0.001	0.001
b_agectd_F	-0.873	0.270	-1.394	-0.466	0.003	0.002
b_histoother_F	-0.189	0.573	-1.338	0.938	0.003	0.002
b_histoother_tx	0.113	0.664	-1.211	1.425	0.003	0.002
b_histoother_y	0.013	0.210	-0.401	0.429	0.001	0.001
b_histoothertx_y	-0.085	0.164	-0.409	0.236	0.000	0.000
b_histosquamous_F	-0.878	0.527	-1.964	0.102	0.003	0.002
b_histosquamous_tx	0.089	0.618	-1.084	1.338	0.003	0.002
b_histosquamous_y	0.226	0.198	-0.170	0.599	0.001	0.001
b_histosquamoustx_y	-0.114	0.156	-0.420	0.190	0.000	0.000
b_sIIIB_F	-0.517	0.492	-1.495	0.421	0.002	0.002
b_sIIIB_tx	-0.771	0.545	-1.841	0.331	0.002	0.002
b_sIIIB_y	0.133	0.157	-0.176	0.443	0.001	0.000
b_sIIIBtx_y	0.169	0.151	-0.126	0.465	0.000	0.000
b_txbinary_y	-0.019	0.322	-0.645	0.614	0.002	0.001
b_wtlossany_F	-1.273	0.566	-2.395	-0.255	0.004	0.003
b_wtlossany_tx	1.494	0.650	0.244	2.778	0.003	0.002
b_wtlossany_y	-0.135	0.202	-0.539	0.246	0.001	0.001
b_wtlossanytx_y	0.161	0.155	-0.140	0.467	0.000	0.000
mu_eGFRunder60	1.251	0.376	0.508	1.981	0.002	0.001
mu_ecogbinary1	-0.838	0.293	-1.418	-0.273	0.002	0.001
mu_tx	3.667	1.051	1.698	5.775	0.008	0.006
alpha0	0.743	0.069	0.608	0.880	0.000	0.000
beta0	-0.276	0.382	-1.010	0.467	0.003	0.002
nu0	-0.061	0.057	-0.169	0.054	0.000	0.000
F_mu	0.482	0.017	0.449	0.514	0.000	0.000
F_sd	0.285	0.026	0.242	0.337	0.000	0.000
b_txbinary_y_marginal	0.010	0.223	-0.409	0.450	0.002	0.001
pmodel[0]	0.272	0.209	0.000	0.682	0.001	0.000
pmodel[1]	0.371	0.242	0.000	0.808	0.001	0.000
pmodel[2]	0.358	0.241	0.000	0.797	0.001	0.001

Table 3: Sampling statistics for parameters of the final bayesian model average. The variables F_mu, F_sd and b_txbinary_y_marginal are not parameters of the model, but are deterministically calculated from the parameters, and are included for reference. The parameters pmodel[0], pmodel[1] and pmodel[2] are the model probabilities for the three selected hyperparameter settings: $\sigma_{F\to t_x} = 2.5$, $\sigma_{F\to y} \in \{0.1, 1.0, 2.5\}$

	ess_mean	ess_sd	ess_bulk	ess_tail	r_hat
b_F_eGFRunder60	50491.0	50491.0	47940.0	57108.0	1.0
b_F_ecogbinary1	20092.0	20092.0	19587.0	31950.0	1.0
b_F_tx	12583.0	12583.0	12328.0	21622.0	1.0
b_F_y	13274.0	12712.0	14190.0	21844.0	1.0
b_Ftx_y	112564.0	79754.0	112564.0	97384.0	1.0
b_agectd_F	11479.0	11479.0	12611.0	19832.0	1.0
b_histoother_F	50821.0	35482.0	53186.0	52046.0	1.0
b_histoother_tx	63335.0	63335.0	63735.0	75460.0	1.0
b_histoother_y	80292.0	62519.0	81810.0	77530.0	1.0
b_histoothertx_y	211765.0	73591.0	211766.0	95551.0	1.0
b_histosquamous_F	38386.0	30373.0	41861.0	41783.0	1.0
b_histosquamous_tx	57878.0	57878.0	59599.0	68582.0	1.0
b_histosquamous_y	41188.0	41188.0	43284.0	54958.0	1.0
b_histosquamoustx_y	196417.0	85781.0	196450.0	94168.0	1.0
b_sIIIB_F	42614.0	24811.0	49835.0	41523.0	1.0
b_sIIIB_tx	64364.0	64364.0	65085.0	73268.0	1.0
b_sIIIB_y	93295.0	90465.0	95162.0	76567.0	1.0
b_sIIIBtx_y	223477.0	115619.0	223486.0	92992.0	1.0
b_txbinary_y	29011.0	29011.0	30389.0	37142.0	1.0
b_wtlossany_F	26035.0	24869.0	28244.0	37178.0	1.0
b_wtlossany_tx	46061.0	46061.0	44526.0	49248.0	1.0
b_wtlossany_y	32376.0	32376.0	33430.0	50794.0	1.0
b_wtlossanytx_y	199570.0	107291.0	199624.0	94483.0	1.0
mu_eGFRunder60	47445.0	45579.0	47018.0	61435.0	1.0
mu_ecogbinary1	22216.0	22216.0	21742.0	36303.0	1.0
mu_tx	16716.0	16716.0	16307.0	28481.0	1.0
alpha0	188062.0	188062.0	190422.0	94694.0	1.0
beta0	17370.0	17370.0	17917.0	29165.0	1.0
nu0	75598.0	75598.0	78518.0	66719.0	1.0
F_mu	87959.0	87959.0	87552.0	85513.0	1.0
F_sd	12815.0	12677.0	13582.0	21807.0	1.0
b_txbinary_y_marginal	15367.0	15367.0	16531.0	23319.0	1.0
pmodel[0]	144220.0	105733.0	143474.0	67157.0	1.0
pmodel[1]	171243.0	134171.0	165659.0	73870.0	1.0
pmodel[2]	137870.0	110311.0	134905.0	72606.0	1.0

E Discussion

E.1 Potential Applications of PROTECT

We now provide two examples where the PROTECT method may be used.

example 1: laryngeal carcinoma For locally advanced unresectable squamous cell laryngeal carcinoma, concurrent chemoradiation is recommended. For patients over 70 years of age or who are unfit for concurrent treatment, only radiotherapy is recommended [13]. RCTs comparing concurrent chemoradiation with radiotherapy alone will be conducted in populations where concurrent chemoradiation is a feasible treatment. Therefore, the RCTs provide no evidence on the treatment effect of concurrent treatment for older and weaker patients. In real-world clinical practice, older and weaker patients may receive the concurrent treatment for example when they have a strong preference for this treatment. Observational studies comparing both treatments will have to deal with the same unobserved confounder as in the NSCLC case: the overall fitness of the patient. To estimate the treatment effect in the older and weaker population, PROTECT can be used to estimate the treatment effect from observational data.

example 2: esophageal cancer The second example concerns stage III squamous cell esophageal cancer. For these patients, neoadjuvant chemoradiation followed by esophagectomy is considered the most effective treatment while definitive chemoradiation without surgery is recommended for patients who are unfit for surgery or refuse surgery [11]. Whether chemoradiation followed by surgery should be preferred for every patient who is fit enough for surgery remains unanswered, as all RCTs are conducted with only patients who are fit for surgery and receive some form of surgical treatment [9, 5], or with patients who are unfit for surgery and never receive surgical treatment [14, 6]. In real-world clinical practice there will be patients who are deemed fit enough for surgery but refuse the surgery due to reasons not related to their prognosis, e.g. personal preference. Retrospective comparisons of chemoradiation with surgery versus chemoradiation alone will arguably be biased due to the same unobserved confounding as in the case of stage III NSCLC: the overall fitness of the patient, this time specifically fitness for surgical treatment. To answer the question whether surgery should be preferred for every patient who is fit enough, the PROTECT method could be applied to to observational data to mitigate the confounding bias.

References

[1] Anne Aupérin et al. "Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non-small-cell lung cancer". en. In: *J. Clin. Oncol.* 28.13 (May 2010), pp. 2181–2190.

[2] Kenneth A. Bollen. "Confirmatory Factor Analysis". en. In: *Structural Equations with Latent Variables.* Section: Seven _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118619179.ch7. John Wiley & Sons, Ltd, 1989, pp. 226–318. ISBN: 978-1-118-61917-9. DOI: `10.1002/9781118619179.ch7`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118619179.ch7` (visited on 12/28/2020).

[3] James Bradbury et al. *JAX: composable transformations of Python+NumPy programs.* 2018. URL: `http://github.com/google/jax`.

[4] Kevin Burke, M. C. Jones, and Angela Noufaily. "A Flexible Parametric Modelling Framework for Survival Analysis". en. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 69.2 (Apr. 2020). Number: 2, pp. 429–457. ISSN: 1467-9876. URL: `http://oro.open.ac.uk/69193/` (visited on 01/11/2021).

[5] Federico Coccolini et al. "Neoadjuvant chemotherapy in advanced gastric and esophago-gastric cancer. Meta-analysis of randomized trials". eng. In: *International Journal of Surgery (London, England)* 51 (Mar. 2018), pp. 120–127. ISSN: 1743-9159. DOI: `10.1016/j.ijsu.2018.01.008`.

[6] Thierry Conroy et al. "Definitive chemoradiotherapy with FOLFOX versus fluorouracil and cisplatin in patients with oesophageal cancer (PRODIGE5/ACCORD17): final results of a randomised, phase 2/3 trial". eng. In: *The Lancet. Oncology* 15.3 (Mar. 2014), pp. 305–314. ISSN: 1474-5488. DOI: `10.1016/S1470-2045(14)70028-2`.

[7] Issa J. Dahabreh, James M. Robins, and Miguel A. Hernán. "Benchmarking Observational Methods by Comparing Randomized Trials and Their Emulations". eng. In: *Epidemiology (Cambridge, Mass.)* 31.5 (Sept. 2020), pp. 614–619. ISSN: 1531-5487. DOI: `10.1097/EDE.0000000000001231`.

[8] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." en. In: *Annals of Statistics* 29.5 (Oct. 2001). Publisher: Institute of Mathematical Statistics, pp. 1189–1232. ISSN: 0090-5364, 2168-8966. DOI: `10.1214/aos/1013203451`. URL: `https://projecteuclid.org/euclid.aos/1013203451` (visited on 12/31/2020).

[9] P. van Hagen et al. "Preoperative chemoradiotherapy for esophageal or junctional cancer". eng. In: *The New England Journal of Medicine* 366.22 (May 2012), pp. 2074–2084. ISSN: 1533-4406. DOI: `10.1056/NEJMoa1112088`.

[10] Miguel A Hernán and James M Robins. *Causal Inference: What If.* en. Boca Raton: Champan & Hall/CRC, 2020.

[11] A. Ajani Jaffer. *NCCN Esophageal and Esophagogastric Junction Cancers, Version 1.2021.* Feb. 2021. URL: `https://www.nccn.org/professionals/physician_gls/pdf/esophageal.pdf` (visited on 02/15/2021).

[12] Christos Louizos et al. "Causal Effect Inference with Deep Latent-Variable Models". In: (May 2017). arXiv: `1705.08821 [stat.ML]`.

[13] J.-P. Machiels et al. "Squamous cell carcinoma of the oral cavity, larynx, oropharynx and hypopharynx: EHNS–ESMO–ESTRO Clinical Practice Guidelines for diagnosis, treatment and follow-up†". English. In: *Annals of Oncology* 31.11 (Nov. 2020). Publisher: Elsevier, pp. 1462–1475. ISSN: 0923-7534, 1569-8041. DOI: `10.1016/j.annonc.2020.07.011`. URL: `https://www.annalsofoncology.org/article/S0923-7534(20)39949-X/abstract` (visited on 02/23/2021).

[14] Bruce D. Minsky et al. "INT 0123 (Radiation Therapy Oncology Group 94-05) phase III trial of combined-modality therapy for esophageal cancer: high-dose versus standard-dose radiation therapy". eng. In: *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 20.5 (Mar. 2002), pp. 1167–1174. ISSN: 0732-183X. DOI: `10.1200/JCO.2002.20.5.1167`.

[15] M M Oken et al. "Toxicity and response criteria of the Eastern Cooperative Oncology Group". en. In: *Am. J. Clin. Oncol.* 5.6 (Dec. 1982), pp. 649–655.

[16] "Causal Diagrams and the Identification of Causal Effects". In: *Causality.* Ed. by Judea Pearl. Cambridge: Cambridge University Press, 2009, pp. 79–80. ISBN: 978-0-521-89560-6. DOI: `10.1017/CBO9780511803161.005`. URL: `https://www.cambridge.org/core/books/causality/causal-diagrams-and-the-identification-of-causal-effects/D9AE074727C3AC9AFE9F0CD4C7A506B5` (visited on 12/28/2020).

[17] Judea Pearl. *Causality*. en. Cambridge University Press, Sept. 2009.

[18] Du Phan, Neeraj Pradhan, and Martin Jankowiak. “Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro”. In: *arXiv preprint arXiv:1912.11554* (2019).

[19] Ricardo Silva and Zoubin Ghahramani. “The Hidden Life of Latent Variables: Bayesian Learning with Mixed Graph Models”. en. In: (), p. 52.

[20] Anders Skrondal and Sophia Rabe-Hesketh. “5.2.5 Empirical identification”. en. In: *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. URL: `https://www.routledge.com/Generalized-Latent-Variable-Modeling-Multilevel-Longitudinal-and-Structural/Skrondal-Rabe-Hesketh/p/book/9781584880004` (visited on 03/03/2021).

[21] Tyler J. VanderWeele. “On the distinction between interaction and effect modification”. eng. In: *Epidemiology (Cambridge, Mass.)* 20.6 (Nov. 2009), pp. 863–871. ISSN: 1531-5487. DOI: `10.1097/EDE.0b013e3181ba333c`.

[22] Aki Vehtari, Andrew Gelman, and Jonah Gabry. “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. en. In: *Statistics and Computing* 27.5 (Sept. 2017), pp. 1413–1432. ISSN: 0960-3174, 1573-1375. DOI: `10.1007/s11222-016-9696-4`. URL: `http://link.springer.com/10.1007/s11222-016-9696-4` (visited on 10/05/2020).

[23] Ian R. White and John B. Carlin. “Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values”. en. In: *Statistics in Medicine* 29.28 (2010). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.3944, pp. 2920–2931. ISSN: 1097-0258. DOI: `10.1002/sim.3944`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3944` (visited on 06/30/2020).

CHAPTER 6

Conditional average treatment effect estimation with treatment offset models

Wouter A.C. van Amsterdam[1,*] and Rajesh Ranganath[2]

[1] Babylon Health, * wouter.vanamsterdam@babylonhealth.com.
[2] Courant Institute of Mathematical Sciences, Center for Data Science, New York University

Abstract

Treatment effect estimates are often available from randomized controlled trials as a single *average treatment effect* for a certain patient population. Estimates of the *conditional average treatment effect* (CATE) are more useful for individualized treatment decision making, but randomized trials are often too small to estimate the CATE. There are several examples in medical literature where the assumption of a known constant *relative* treatment effect (e.g. an odds-ratio) is used to estimate CATE models from large observational datasets. One approach to estimating these CATE models is by using the relative treatment effect as an *offset*, while estimating the covariate-specific baseline risk. Whether this is a valid approach in the presence of unobserved confounding is unknown.

We demonstrate for a simple example that offset models do not recover the true CATE in the presence of unobserved confounding. We then explore the magnitude of this bias in numerical experiments. For virtually all plausible confounding magnitudes, estimating the CATE using offset models is more accurate than assuming a single absolute treatment effect whenever there is sufficient variation in the baseline risk. Next, we observe that the odds-ratios reported in randomized controlled trials are not the odds-ratios that are needed in offset models because trials often report the *marginal* odds-ratio. We introduce a constraint to better use marginal odds-ratios from randomized controlled trials and find that the newly introduced constrained offset models have lower bias than standard offset models. Finally, we highlight directions for future research for exploiting the assumption of a constant relative treatment effect with offset models.

1 Introduction

Weighing potential benefits and harms of treatment requires knowing the treatment effect, which is the change in probability of an outcome between different treatments. The gold standard for estimating treatment effects are randomized control trials (RCT). RCTs often report the efficacy of treatments on a relative scale using for example the odds-ratio or hazard-ratio for the entire population (e.g. Furie et al. (2020); Lean et al. (2018)). Though reported on a relative scale, the effect reported still corresponds to a single average effect, i.e., absolute change in probability of an outcome, for the whole population. Ideally, the change in probability of an outcome, rather than being known on average for a population, would be tailored to the characteristics of a patient to produce the conditional average treatment effect (CATE). To turn the population-level relative effect into a CATE estimate, several previous studies on breast cancer and cardiovascular disease used the assumption of a *constant relative treatment effect* to develop CATE models from observational data (Candido dos Reis et al., 2017; Ravdin et al., 2001; Alaa et al., 2021; Xu et al., 2021). We call these constant-relative CATE (CR-CATE) models. This assumption can be better contextualized by imagining studying the effect of a new type of medication, here the two “treatments” being compared are the untreated or baseline regime and that same regime plus this new medication. The assumption of a constant relative treatment effect does not preclude non-constant CATEs because even with a constant relative treatment effect, the treatment can have a varying effect on an absolute risk scale depending on the baseline risk of a patient. For instance, assume that a new cholesterol lowering drug reduces the risk of cardiovascular death within the next 10 years with an odds-ratio of 0.5. A 60-year-old male smoker with hypertension and raised cholesterol has a baseline risk of cardiovascular death of 40% and should expect a reduction in risk of 15% points. A 50-year-old female without hypertension has a baseline risk of under 1% and will have a less than 0.5% points reduction in risk. Given these widely different effects on an absolute probability scale, one may recommend the new cholesterol lowering drug to the 60-year-old male but not the 50-year-old female.

When estimating CR-CATE models from observational data where the treatment of interest was available, one approach is to use the constant relative treatment effect as an *offset* term, while estimating the baseline risk. By combining the estimated baseline risk model and the fixed relative treatment effect, these models estimate the absolute outcome probability under treatment or no treatment. Some CR-CATE were found to be accurate in observational validation studies, on the basis of which treatment guidelines acknowledged a place for them in clinical decision making (Cardoso et al., 2019; Gradishar, 2021). However, due to confounding, the baseline risk cannot be estimated from an observational dataset where some patients were treated and others were not (Groenwold et al., 2016; van Geloven et al., 2020).

Because CR-CATE models target interventional distributions but were developed from observational data it is implicitly assumed that using the constant relative treatment effect assumption is sufficient for controlling for any unobserved confounding. At first glance this implicit assumption may seem plausible as the constant relative treatment effect is not estimated from the observational data but is plugged in from prior RCT estimates. However, whether the assumption is correct has not been discussed or verified.

In this work we evaluate the validity of the assumption that a known constant odds-ratio for treatment allows for CATE estimation in the presence of unobserved confounding using the known odds-ratio as an offset term. We show that a known odds-ratio used as an offset is not sufficient for estimating CATEs. In spite of that, we find in numerical experiments the bias was low enough that using offset models still leads to better estimation of CATEs compared with the baseline of assuming a single risk difference for all patients. Finally, we observe that the odds-ratios reported in RCTs are not the odds-ratios that are needed for the offset method because RCTs generally report estimates of the *marginal* odds-ratio, whereas the offset method requires the *conditional* odds-ratio. We introduce a constraint to the offset method that restricts the offset models based on the *marginal* odds-ratio from the RCT, and find empirically that these constrained offset models have lower bias.

2 Methods

We consider models that estimate the absolute difference in probability of a binary outcome y under two possible treatments $t_x \in \{0, 1\}$ conditional on a possibly multi-dimensional pre-treatment covariate vector $\boldsymbol{x}$. This is the conditional average treatment effect (CATE), conditional on $\boldsymbol{x}$. Treatment $t_x = 0$ is assumed to be the baseline treatment (or no treatment depending on the clinical context) and $t_x = 1$ is the treatment of interest. Using Pearl’s do-operator to indicate *intervening* on treatment, the CATE is defined as:

$$\text{CATE}(\boldsymbol{x}) := p(y = 1|\text{do}(t_x = 1), \boldsymbol{x}) - p(y = 1|\text{do}(t_x = 0), \boldsymbol{x})$$

CR-CATE approaches assume that the odds-ratio for treatment is constant for the entire population. Odds are defined relative to a probability π as $\text{odds}(\pi) = \frac{\pi}{1-\pi}$. The odds-ratio of two probabilities π_0, π_1 is defined as $\text{OR}(\pi_1, \pi_0) := \text{odds}(\pi_1)/\text{odds}(\pi_0)$. Writing $\pi_{t'_x}(\boldsymbol{x'}) = p(y = 1|\text{do}(t_x = t'_x), \boldsymbol{x} = \boldsymbol{x'})$, the assumption that the odds-ratio for treatment is constant implies that for any two possible values $\boldsymbol{x}', \boldsymbol{x}''$ for $\boldsymbol{x}$, $\text{OR}(\pi_1(\boldsymbol{x}'), \pi_0(\boldsymbol{x}')) = \text{OR}(\pi_1(\boldsymbol{x}''), \pi_0(\boldsymbol{x}''))$. Or equivalently, the log odds of the interventional distributions differ by a constant. Introducing $\eta(t_x, \boldsymbol{x})$ as the log odds of $\pi_{t_x}(x)$, the assumption implies that for each $t_x, \boldsymbol{x}$:

$$\eta(t_x, \boldsymbol{x}) = \beta_0(\boldsymbol{x}) + \beta^*_{t_x} t_x \tag{1}$$

Here $\beta_0(\boldsymbol{x})$ is the log odds of the interventional distribution with $\text{do}(t_x = 0)$ (i.e. the baseline risk) as a function of $\boldsymbol{x}$, and $\beta^*_{t_x}$ is the log odds-ratio for treatment, assumed to be constant for all $\boldsymbol{x}$. As we will explain later in Section 2.2, $\beta^*_{t_x}$ depends on the choice for the covariate $\boldsymbol{x}$. For the moment we will assume that an estimate of $\beta^*_{t_x}$ for the chosen $\boldsymbol{x}$ is available from prior RCTs. Later in Section 2.2 we discuss the more realistic setting where this is not the case. Denote $\sigma(x) = (1 + e^{-x})^{-1}$ the sigmoid function and $\sigma^{-1}(\pi) = \log \text{odds}(\pi); 0 < \pi < 1$ its inverse, we can now write the CATE in terms of η:

$$\text{CATE}(\boldsymbol{x}) = \sigma(\eta(1, \boldsymbol{x})) - \sigma(\eta(0, \boldsymbol{x}))$$

If $\beta^*_{t_x}$ is given a priori, models for η of the form of Equation 1 can be estimated with likelihood based approaches by specifying a parametric model for $\hat{\beta}_0(\boldsymbol{x}) = f(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$ where $f : \mathcal{X} \times \Theta \to \mathbb{R}$ is from a family of functions of $\boldsymbol{x} \in \mathcal{X}$ indexed by parameter vector $\boldsymbol{\theta} \in \Theta$. The full model is then given by

$$\hat{\eta}(t_x, \boldsymbol{x}) = f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) + \beta^*_{t_x} t_x \tag{2}$$

In the case of logistic regression, $f(\boldsymbol{x}; \boldsymbol{\theta}) = \theta_0 + \boldsymbol{\theta}_{\boldsymbol{x}} \boldsymbol{x}$. A fixed term in a model that is not estimated from data is called an *offset* term (Watson, 2007). We therefore refer to models of the from of Equation 2 where $\beta^*_{t_x}$ is constant as *treatment offset models* or *offset models* for short. Offset models are a subclass of CR-CATE models.

2.1 Identification of the conditional average treatment effect

We assume we are given data from an observational distribution compatible with the Acyclic Mixed Directed Graph (AMDG) with observed multi-dimensional covariate vector $\boldsymbol{x}$ and unobserved confounder $\boldsymbol{u}$ presented in Figure 1. The AMDG is quite general in that it allows for unobserved confounding between the pairs of variables $[\{\boldsymbol{u}, \boldsymbol{x}\}, \{\boldsymbol{x}, t_x\}, \{\boldsymbol{x}, y\}]$. It is assumed that $\boldsymbol{x}$ is a non-descendant of t_x, y as implied by the assumption that $\boldsymbol{x}$ is a pre-treatment variable that may be useful for individual treatment decisions.

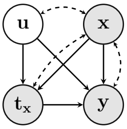


Figure 1: Acyclic Mixed Directed Graph with observed nodes $t_x, \boldsymbol{x}, y$ and unobserved confounder $\boldsymbol{u}$. Double arrows indicate the presence of a confounder, meaning that $u \leftarrow z_{ux} \rightarrow x \iff u \longleftrightarrow x$.

To prove that the CATE is identified it is sufficient to prove that the interventional distribution $\pi_{t'_x}(\boldsymbol{x}')$ is identified for all $t'_x, \boldsymbol{x'}$. Due to the unobserved confounder $\boldsymbol{u}$ in the AMDG, the interventional distribtution is not identifiable from observational data without additional assumptions. The question is whether the assumption of a known constant log odds-ratio as stated in Equation 1 is sufficient for $\boldsymbol{x}$-conditional causal effect identification from observational data when $\boldsymbol{u}$ is not observed. The known constant odds-ratio assumption implies that the query is identified when the baseline risk $\sigma(\beta_0(\boldsymbol{x}))$ is identified, as $\pi_0(\boldsymbol{x}') = \sigma(\beta_0(\boldsymbol{x}'))$ and $\pi_1(\boldsymbol{x}') = \sigma(\beta_0(\boldsymbol{x}') + \beta^*_{t_x})$. We now prove with a simple counter example that offset models do not estimate the ground truth interventional distribution.

2.1.1 Example 1: Offset models do not estimate the interventional distribution

A simple example compatible with Equation 1 and the AMDG Figure 1 is where $\boldsymbol{u}$ is binary and $\beta_0(\boldsymbol{x}) = \beta_0^*$ for all values of $\boldsymbol{x}$, meaning that there is no variation in the baseline risk. Denoting $\mathcal{B}$ as the Bernoulli distribution, $p_u = p(u = 1)$ and $\pi_{t_x u} = p(y = 1|t_x, u)$, then the data-generating mechanism for this example is:

$$u \sim \mathcal{B}(p_u), t_x \sim \mathcal{B}(p(t_x = 1|u = u)), y \sim \mathcal{B}(\pi_{t_x u}) \tag{3}$$

Despite its simplicity this example is conceptually important for all cases with binary treatment and discrete $\boldsymbol{x}$ as a) when the treatment is binary, any arbitrary confounder can be modeled as a single binary variable while maintaining the same observational and interventional distributions (Ilse et al., 2022); and b) in the limit of infinite data, stratifying the population for each value of $\boldsymbol{x}$ and estimating β_0 in each of the strata is equivalent to non-parametric estimation of $\beta_0(\boldsymbol{x})$ in Equation 2 when u is binary and $\boldsymbol{x}$ is discrete.

Let $l(y, \hat{\pi})$ denote the Bernoulli log-likelihood of outcome y conditional on estimated probability $\hat{\pi}$. We derive a closed-form expression for the expected log-likelihood depending on the single parameter of the offset model β_0: $L(\beta_0) = E_{p_{\text{obs}}(y,t_x,u)}[l(y, \hat{\pi}(t_x, u, \beta_0)]$ in the Appendix A.1. Taking the derivative with respect to β_0 and plugging in the ground truth value for β_0^* we find the following expression:

$$\begin{aligned}\frac{\partial L}{\partial \beta_0}\left(\beta_0 = \beta_0^*\right) = p_u(1 - p_u)\big[&(\pi_{01} - \pi_{00})\left(p(t_x = 0|u = 1) - p(t_x = 0|u = 0)\right) + \\ &(\pi_{11} - \pi_{10})\left(p(t_x = 1|u = 1) - p(t_x = 1|u = 0)\right)\big]\end{aligned}$$

In general this expression is non-zero, meaning that the ground truth solution β_0^* is not a stationary point of the expected log-likelihood. This proves that the offset method does not recover the true baseline risk in the presence of confounding. In the case of no confounding when $\pi_{t_x 0} = \pi_{t_x 1}$ or $p(t_x|u = 1) = p(t_x|u = 0)$, the derivative is zero at β_0^* meaning β_0^* is a stationary point of the expected log-likelihood. The question is now how important this bias is and whether offset models can be used if the constant odds-ratio assumption is tenable, or if offset models should be avoided altogether. We study this later in with numerical experiments in Section 3.1.

2.2 Collapsibility

An important consideration for offset models is the difference between the *marginal* odds-ratio and the *conditional* odds-ratio. In a sufficiently large RCT where treatment t_x is randomized and binary covariate x is observed, two different models may be estimated: one that does not condition on x and estimates the *marginal* log odds-ratio γ_{t_x}:

$$p(y = 1|\text{do}(t_x = t_x')) = \sigma(\gamma_0 + \gamma_{t_x} t_x')$$

and one that estimates the *conditional* log odds-ratio β_{t_x}:

$$p(y = 1|\text{do}(t_x = t_x'), x = x') = \sigma\left(\beta_0 + \beta_{t_x} t_x' + \beta_x x'\right)$$

Note that the model with the conditional odds-ratio does not include an interaction term between t_x and x, as implied by the constant odds-ratio assumption. In contrast with linear regression, in general $\beta_{t_x} \neq \gamma_{t_x}$. This means that the odds-ratio for treatment is different if the model conditions on the covariate x or not. This property of the odds-ratio is called *non-collapsibility* (Greenland et al., 1999; Burgess, 2017). To illustrate non-collapsibility, consider the extreme example with binary covariate x where $\pi_{\{0,1\}}(x = 0) = \{0.01, 0.02\}$ and $\pi_{\{0,1\}}(x = 1) = \{0.98, 0.99\}$. For both $x \in \{0, 1\}$, the *x-conditional* log odds-ratio $\beta_{t_x} \approx \log(2.0)$. However, when grouping patients with different values of x together we see that, assuming $p(x = 1) = 0.5$, $p(y = 1|\text{do}(t_x = \{0, 1\}) = \{0.495, 0.505\}$, thus the *marginal* log odds-ratio $\gamma_x \approx \log(1.0)$. In most RCTs the *marginal* log odds-ratio γ_{t_x} is estimated. When $\gamma_{t_x} \neq \beta_{t_x}$ the trials do not provide the information required to use the offset method as defined in Equation 2. The stronger the t_x-conditional association between x and y, the greater the difference between γ_{t_x} and β_{t_x} (Hauck et al., 1991). For an illustration, see Appendix A.2. This is an important drawback as at the same time, a stronger association between x and y conditional on t_x results in more variation in the baseline risk and thus more variation in the x-conditional treatment effect. So in the situation where offset models have more potential added value (when x-conditional treatment effects differ substantially), the estimate of the marginal log odds-ratio γ_{t_x} from RCTs becomes a less accurate approximation of the β_{t_x} needed for the offset model.

Having defined β_{t_x} and γ_{t_x} we can now refine the assumptions underlying binary treatment offset models stating that 1) the *conditional* odds-ratio β_{t_x} does not depend on $\boldsymbol{x}$ as in Equation 1, and 2) the *marginal* log odds-ratio γ_{t_x} is known from RCTs.

2.2.1 Using the marginal odds-ratio as a constraint

Instead of using $\gamma^*_{t_x}$ in the place of $\beta^*_{t_x}$ in an offset model we now propose a new approach for using knowledge of $\gamma^*_{t_x}$. We do not have knowledge of $\beta^*_{t_x}$ in Equation 2, instead, we estimate β_{t_x} as a parameter alongside $\boldsymbol{\theta}$ from the observational data. Given an estimate of the parameters $(\hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}})$ we can estimate what the marginal odds-ratio $\gamma_{t_x}(\hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}})$ would have been if an RCT had been conducted in the same patient population. We call this $\gamma_{t_x}(\hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}})$ the *implied* marginal log odds-ratio. Specifically, for each observed value $\boldsymbol{x}' \in \mathcal{X}$, we can calculate the two estimated interventional outcome distributions:

$$p(y = 1|\text{do}(t_x = 0), \boldsymbol{x} = \boldsymbol{x}'; \hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}}) = \sigma(f(\boldsymbol{x}'; \hat{\boldsymbol{\theta}})) \tag{4}$$

$$p(y = 1|\text{do}(t_x = 1), \boldsymbol{x} = \boldsymbol{x}'; \hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}}) = \sigma(f(\boldsymbol{x}'; \hat{\boldsymbol{\theta}}) + \hat{\beta}_{t_x}) \tag{5}$$

The joint distribution and data generating mechanism for an RCT in this population with (unknown) distribution $p_{\text{RCT}}(\boldsymbol{x}) = p(\boldsymbol{x})$ of $\boldsymbol{x}$ and treatment probability $p_{\text{RCT}}(t_x = 1)$ are:

$$p_{\text{RCT}}(y, t_x, \boldsymbol{x}) = p(y|t_x, \boldsymbol{x}) p_{\text{RCT}}(t_x|\boldsymbol{x}) p(\boldsymbol{x}) = p(y|t_x, \boldsymbol{x}) p_{\text{RCT}}(t_x) p(\boldsymbol{x}) \tag{6}$$

$$\boldsymbol{x} \sim p(\boldsymbol{x}), t_x \sim \mathcal{B}(p_{\text{RCT}}(t_x = 1)), y \sim \mathcal{B}(\sigma(f(\boldsymbol{x}'; \hat{\boldsymbol{\theta}}) + \hat{\beta}_{t_x} t_x)) \tag{7}$$

The maximum likelihood estimate of the marginal log odds-ratio γ_{t_x} in the data generating mechanism of this hypothetical RCT is:

$$\gamma_{t_x} = \sigma^{-1}(p_{\text{RCT}}(y = 1|t_x = 1)) - \sigma^{-1}(p_{\text{RCT}}(y = 1|t_x = 0))$$

We can use Equations 4 and 5 to calculate $p_{\text{RCT}}(y = 1|t_x = t'_x; \hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}})$ by averaging over $\boldsymbol{x}$. Given the distribution in 6 we can do this averaging using $p(\boldsymbol{x})$ because $\boldsymbol{x}$ and t_x are marginally independent in the hypothetical RCT. This leads to:

$$p_{\text{RCT}}(y = 1|t_x = t'_x; \hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}}) = E_{\boldsymbol{x}' \sim p(\boldsymbol{x})} \sigma(f(\boldsymbol{x}'; \hat{\boldsymbol{\theta}}) + \hat{\beta}_{t_x} t'_x)$$

Because we generally do not know $p(\boldsymbol{x})$ we replace the expectation with the mean over observed values $\boldsymbol{x}_i$ of $\boldsymbol{x}$ from the empirical distribution and arrive at our estimate of the implied marginal odds-ratio:

$$\gamma_{t_x}(\hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}}) = \sigma^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} \sigma(f(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}) + \hat{\beta}_{t_x})\right) - \sigma^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} \sigma(f(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}))\right) \tag{8}$$

Given an estimate of the marginal odds-ratio $\gamma^*_{t_x}$ from RCTs, we can now formulate a new objective that includes both the likelihood of the observed data and a constraint defined by the known versus implied marginal odds-ratio. Denote the Bernoulli log-likelihood of an individual observation as $l(y_i, x_i; \hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}})$ and $L(\hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} l(y_i, x_i; \hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}})$ the total log-likelihood of the observed data. We formulate the following Lagrangian:

$$\mathcal{L}(\hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}}) = L(\hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}}) + \lambda\left(\gamma_{t_x}(\hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}}) - \gamma^*_{t_x}\right) \tag{9}$$

The optimal set of parameters maximizes the likelihood of the observed data while adhering to the constraint on the implied marginal odds-ratio. Finding these parameters given data can be done with constrained optimization algorithms, for example an Augmented Lagrangian Algorithm (Madsen et al., 2004).

2.3 Metric

CATE models estimate the difference in outcome probability under hypothetical interventions on treatment conditional on covariates. A common metric for CATE estimation is the root-mean-squared error of the predicted difference in outcome probability versus the actual difference in outcome probability, also known as the "Precision in Treatment Effect Heterogeneity" (PEHE, Hill (2011)). If $\pi_1(\boldsymbol{x}), \pi_0(\boldsymbol{x})$ denote the interventional distributions, and $\hat{\pi}_1(\boldsymbol{x}), \hat{\pi}_0(\boldsymbol{x})$ the estimated interventional distributions, the PEHE is calculated as:

$$\text{PEHE} = \sqrt{\frac{1}{N}\sum_{i}^{N}\left((\pi_1(\boldsymbol{x}_i) - \pi_0(\boldsymbol{x}_i)) - (\hat{\pi}_1(\boldsymbol{x}_i) - \hat{\pi}_0(\boldsymbol{x}_i))\right)^2}$$

CATE models are generally motivated to enable more individualized treatment decisions as opposed to using a single average treatment effect estimate for all patients. This means that the baseline for CATE models is using a single average treatment effect on the absolute probability scale for all patients (ATE-baseline).

3 Experiments

We evaluate the amount of bias when estimating CATE models from observational data using the offset method in the presence of unobserved confounding with numerical experiments. First, we investigate Example 1 (Equation 3) and find that the bias of offset models is small even for large confounding magnitudes. Finally we study in what situations offset models have better PEHE than the ATE-baseline when there are measured covariates, comparing different offset model variants.

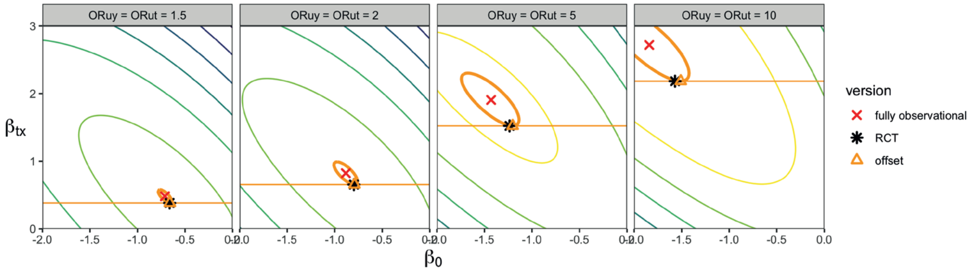


Figure 2: Solutions for different methods on Example 1 with different amounts of confounding, indexed by $\text{OR}_{ut} = \text{OR}_{uy}$ the odds-ratios from confounder u to treatment t_x and outcome y respectively. The contour lines indicate solutions with the same log-likelihood of the observational data. As visualized with the horizontal line, the offset method finds the β_0 that maximizes the observational likelihood while keeping $\beta_{t_x} = \beta^*_{t_x}$. Fully observational: estimate β_0 and β_{t_x} from observational data, RCT: ground truth values of β_0 and β_{t_x}, offset: offset method.

3.1 Example 1 examined

To evaluate the amount of bias in offset models in Example 1, we parameterize the magnitude of confounding using log odds-ratios $\beta_{u\to t_x}, \beta_{u\to y}$ so that $p(t_x = 1|u) = \sigma(\frac{1}{2}\beta_{u\to t_x}(2u-1))$ and $p(y = 1|t_x, u) = \sigma(\frac{1}{2}(\beta_{t_x}(2t_x - 1) + \beta_{u\to y}(2u-1)))$. Note that because there is no variation in baseline risk $\gamma^*_{t_x} = \beta^*_{t_x}$. We plot different solutions and the log-likelihood contours for different values of $\beta_{u\to t_x} = \beta_{u\to y}$ in Figure 2, setting $\beta_{t_x} = 1$ and $p_u = 0.5$. Even in extreme cases of confounding when $\beta_{u\to t_x} = \beta_{u\to y} = \log 10$, the offset solution is close to the ground truth, while the observational estimate becomes more and more biased. This indicates that when $\beta^*_{t_x}$ is known, the bias in offset models induced by the unobserved confounder u is small.

3.2 Numerical experiments with a binary covariate

The bias induced by the confounding in Example 1 seems minor even for extreme magnitudes of confounding when $\beta^*_{t_x}$ is known. However, a more important metric is whether the PEHE of the offset model is better than that of the ATE-baseline when $\gamma^*_{t_x}$ is known instead of $\beta^*_{t_x}$ and there is variation in the baseline risk, which means that $\gamma^*_{t_x} \neq \beta^*_{t_x}$. To investigate this, we extend the example by introducing a marginally independent binary covariate x with non-zero effect on the outcome. The updated data generating mechanism is:

$$u \sim \mathcal{B}(p_u), t_x \sim \mathcal{B}(p(t_x = 1|u = u)), x \sim \mathcal{B}(p_x), y \sim \mathcal{B}(\pi_{t_x x u})$$

where

$$\pi_{t_x x u} = p(y = 1|t_x, x, u) = \sigma(\frac{1}{2}(\beta_{t_x}(2t_x - 1) + \beta_x(2x - 1) + \beta_{u \to y}(2u - 1))) \tag{10}$$

For different values of $\beta_x, \beta_{u \to t_x}, \beta_{u \to y}$ in Equation 10 we calculated the PEHE of the ATE-baseline. We contrast this PEHE with 5 different approaches. As we are investigating the amount of bias due to unobserved confounding, the reference is (1) a logistic regression model based on data where there is no confounding as in RCTs, with $p_{\text{rct}}(t_x = 1|u = 0) = p_{\text{rct}}(t_x = 1|u = 1) = 0.5$, but the rest of the data generating mechanism remains the same (RCT). We then compare 4 different approaches using observational data: (2) a logistic regression model where $\beta_0, \beta_{t_x}, \beta_x$ are estimated from the observational data (full). (3) An offset model where β_0, β_x are estimated while plugging in the ground truth $\beta^*_{t_x}$ as obtained by the RCT in model (1) (conditional). (4) An offset model where the *marginal* $\gamma^*_{t_x}$ is available from RCTs and is used as an offset in place of β_{t_x} (marginal). (5) An offset model where the *implied marginal* $\gamma_{t_x}(\hat{\beta}_{t_x}, \hat{\boldsymbol{\theta}})$ is constrained to be $\gamma^*_{t_x}$ as in Equation 9 (constrained). For these experiments we set $\beta_{u \to t_x} = \beta_{u \to y} = \beta_u$ to four different values and varied β_x, keeping $\beta_{t_x} = 1$ and $p_u = p_x = 0.5$. As for these experiments the expected log-likelihood is available in closed-form, we optimize the expected log-likelihood directly instead of generating random samples. We implemented the constrained offset model using a gradient-based augmented Lagrangian optimizer implemented in the R package ‘alabama‘. The

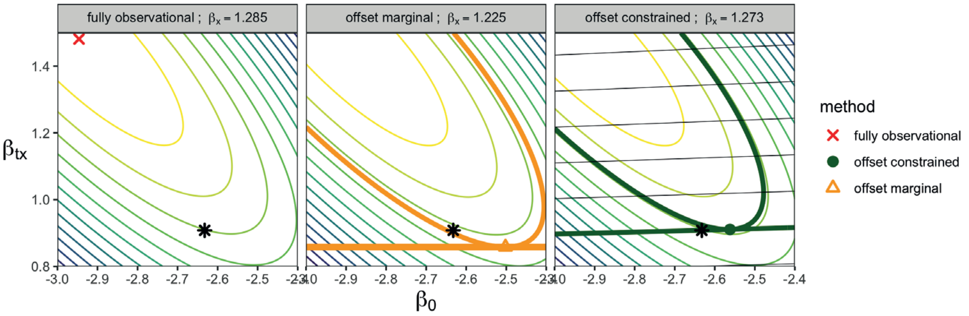


Figure 3: Three different approaches to estimating a model for data with a binary covariate x. Each image is a hyperplane of the parameter cube $(\beta_0, \beta_{t_x}, \beta_x)$ dissected at a specific value of β_x corresponding to the solution of the respective method. The ground truth solution, indicated with the black asterix lies in none of the shown hyperplanes. The contour lines indicate solutions with the same log-likelihood of the observational data. In the *marginal* offset method, the solution for (β_0, β_x) maximizes the log-likelihood on the line with $\beta_{t_x} = \gamma^*_{t_x}$, as indicated with the orange horizontal line in the second plot. Because of non-collapsibility, this is a suboptimal solution as $\gamma^*_{t_x} \neq \beta^*_{t_x}$. In the *constrained* offset method (third plot), reference lines are added that are dissections of level sets defined by equal values of the constraint on the *implied* marginal odds-ratio $\gamma_{t_x}(\beta_0, \beta_{t_x}, \beta_x) - \gamma^*_{t_x}$. Here, the solution $(\beta_0, \beta_{t_x}, \beta_x)$ maximizes the log-likelihood on the level set defined by $\gamma_{t_x}(\beta_0, \beta_{t_x}, \beta_x) - \gamma^*_{t_x} = 0$, which is a saddlepoint of the Lagrangian as formulated in Equation 9.

code to replicate these results is available at www.github.com/vanamsterdam/binaryoffsetmodels. The constrained offset method applied to this setting is illustrated in Figure 3 and constrasted with the fully observational baseline and the marginal offset method.

As a first observation from the results shown in Figure 4, whenever the baseline risk varies with x, the ATE-baseline has sub optimal PEHE. Also, the PEHE of the fully observational logistic regression model becomes worse than the ATE-baseline for higher magnitudes of confounding. Whenever the *estimated* $\widehat{\mathrm{OR}}_x > 1$, offset models are better than the ATE-baseline with respect to PEHE. For larger magnitudes of $\widehat{\mathrm{OR}}_x$ the performance of the *marginal* offset model degrades because the issue of non-collapsibility becomes more pronounced. Of note, the *constrained* offset model is always better than the ATE-baseline whenever $\widehat{\mathrm{OR}}_x > 1$, and always better than the fully observational baseline. Finally, we observe that even the logistic regression model estimated from RCT data has non-zero PEHE which increases when the confounding increases. The reason for this not confounding but parametric form bias. The data were generated according to a simple logistic regression setup, linear in t_x, x, u. When fitting a logistic model conditional on t_x, x in this data, marginalizing out u, simple logistic regression is no longer sufficient. Specifically, the model now requires an added interaction term between t_x and x to be unbiased.

We further expanded this example with numerical experiments where x and u are no longer marginally independent. Details of these experiments are described in the Appendix A.3. Except in some extreme settings when there is very little variation in baseline risk, the constrained offset models have better PEHE than the ATE-baseline. Overall, constrained offset models perform most stable across all settings and have better PEHE than the fully observational baseline and the marginal offset models.

4 Discussion

We evaluated whether the offset method provides valid CR-CATE models for binary outcomes in the presence of unobserved confounding. Though not exact, offset models still have better PEHE than the baseline of using the average treatment effect for all patients even for large confounding magnitudes. In our numerical experiments, this holds even if an estimate of the *marginal* odds-ratio is used from randomized trials instead of the *conditional* odds-ratio. We introduced a new way of using estimates of the *marginal* odds-ratio to address the issue of non-collapsibility of the odds-ratio and find that it gives the best performance overall in terms of PEHE.

An important question for practical applications is when it is valid to assume that the relative treatment effect is indeed constant. There is some evidence from meta-analyses that treatment effect estimates on a relative scale are more stable across different RCTs than treatment effects on an absolute scale (Engels et al., 2000; Sterne and Egger, 2001). However, in some settings there may clear indications for differences in treatment effect on a relative scale. This could hold for example for therapies whose mechanism of action depends on certain genetic mutations. If this is the case and the difference in relative treatment effect is known, this difference could be accounted for accordingly in offset models.

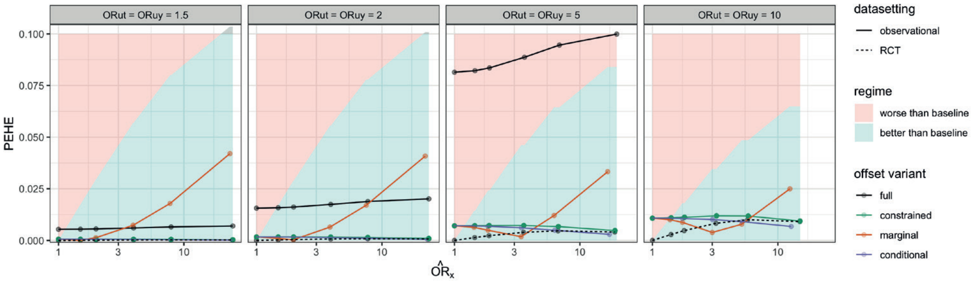


Figure 4: PEHEs for different strategies, indexed by $\mathrm{OR}_{ut} = \mathrm{OR}_{uy}$, the odds-ratios from confounder u to treatment t_x and outcome y respectively. The shaded areas indicate whether the chosen approach improves upon the ATE-baseline of assuming a single predicted difference in outcome for all patients. In the right most plot the fully observational baseline has higher PEHE than the maximum y-value of the plot.

Recent work has studied combining observational data and data from randomized trials for CATE estimation (Rosenman et al., 2020; Ilse et al., 2022). Under relatively mild assumptions, estimates from combined datasets yield more efficient estimates of CATEs than using RCT data alone. However, these methods require access to the individual-patient data from the RCT, whereas offset methods only rely on a single effect estimate from RCTs. Gaining access to individual-patient data from RCTs is often challenging due to data-access restrictions.

A limitation of our work is the relatively restricted set of experiments. Future work could experiment with higher dimensional, mixed-type covariates and different functional relationships between the variables. In higher dimensions, the constraint on the implied marginal odds-ratio restricts a lower fraction of the degrees of freedom. It is unknown whether the constraint will effectively reduce confounding bias in higher dimensions. One potentential solution for this would be to first learn a scalar function from all covariates, for example with a fully observational model or a marginal offset model. The constrained offset method can then be applied using this scalar as the single covariate.

Future work could extend our experiments to relative treatment effect estimates in the form of hazard-ratios, or to the setting of time-varying confounding. Furthermore, Bayesian extensions of our constrained offset model can be investigated to account for uncertainty in marginal odds-ratio estimates from RCTs. Finally, finite-sample characteristics of our estimator for the implied marginal odds-ratio in terms of bias and variance could be studied further. We leave these extensions for future work.

In conclusion, we find that offset models do not correctly estimate CATEs in the presence of unobserved confounding. However, from our experiments it may still be justified to use offset models in practice as they often have better PEHE than the ATE-baseline. The newly introduced constraint on the implied marginal odds-ratio improved the PEHE even more. Further extensions of the offset method for CR-CATE models remain for future work.

Acknowledgments

We are very thankful to Nan van Geloven for providing feedback on an earlier version of this manuscript. The majority of this work was done whileWA was unemployed. WA is now employed by Babylon Health Inc., London, United Kingdom. RR was partly supported by NIH/NHLBI Award R01HL148248, and by NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science. The code to reproduce all experiments will be made publicly available at www.github.com/vanamsterdam/binaryoffsetmodels.

References

Alaa, A. M., Gurdasani, D., Harris, A. L., Rashbass, J., and van der Schaar, M. (2021). Machine learning to guide the use of adjuvant therapies for breast cancer. *Nature Machine Intelligence*, 3(8):716–726. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 8 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Breast cancer;Prognosis Subject_term_id: breast-cancer;prognosis.

Burgess, S. (2017). Estimating and contextualizing the attenuation of odds ratios due to non collapsibility. *Communications in Statistics - Theory and Methods*, 46(2):786–804. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610926.2015.1006778.

Candido dos Reis, F. J., Wishart, G. C., Dicks, E. M., Greenberg, D., Rashbass, J., Schmidt, M. K., van den Broek, A. J., Ellis, I. O., Green, A., Rakha, E., Maishman, T., Eccles, D. M., and Pharoah, P. D. P. (2017). An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Research*, 19(1):58. 80 citations (Crossref) [2021-08-06].

Cardoso, F., Kyriakides, S., Ohno, S., Penault-Llorca, F., Poortmans, P., Rubio, I., Zackrisson, S., and Senkus, E. (2019). Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 30(8):1194–1220.

Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I., and Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine*, 19(13):1707–1728.

Furie, R., Rovin, B. H., Houssiau, F., Malvar, A., Teng, Y. K. O., Contreras, G., Amoura, Z., Yu, X., Mok, C.-C., Santiago, M. B., Saxena, A., Green, Y., Ji, B., Kleoudis, C., Burriss, S. W., Barnett, C., and Roth, D. A. (2020). Two-Year, Randomized, Controlled Trial of Belimumab in Lupus Nephritis. *The New England Journal of Medicine*, 383(12):1117–1128.

Gradishar, W. J. (2021). NCCN Breast Cancer Guideline, Version 5.2021.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1):29–46. Publisher: Institute of Mathematical Statistics.

Groenwold, R. H. H., Moons, K. G. M., Pajouheshnia, R., Altman, D. G., Collins, G. S., Debray, T. P. A., Reitsma, J. B., Riley, R. D., and Peelen, L. M. (2016). Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *Journal of Clinical Epidemiology*, 78:90–100.

Hauck, W. W., Neuhaus, J. M., Kalbfleisch, J. D., and Anderson, S. (1991). A consequence of omitted covariates when estimating odds ratios. *Journal of Clinical Epidemiology*, 44(1):77–81.

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Ilse, M., Forré, P., Welling, M., and Mooij, J. M. (2022). Combining Interventional and Observational Data Using Causal Reductions. *arXiv:2103.04786 [cs, stat]*. arXiv: 2103.04786.

Lean, M. E., Leslie, W. S., Barnes, A. C., Brosnahan, N., Thom, G., McCombie, L., Peters, C., Zhyzhneuskaya, S., Al-Mrabeh, A., Hollingsworth, K. G., Rodrigues, A. M., Rehackova, L., Adamson, A. J., Sniehotta, F. F., Mathers, J. C., Ross, H. M., McIlvenna, Y., Stefanetti, R., Trenell, M., Welsh, P., Kean, S., Ford, I., McConnachie, A., Sattar, N., and Taylor, R. (2018). Primary care-led weight management for remission of type 2 diabetes (DiRECT): an open-label, cluster-randomised trial. *Lancet (London, England)*, 391(10120):541–551.

Madsen, K., Nielsen, H., and Tingleff, O. (2004). *Optimization With Constraints*. IMM, Technical University of Denmark.

Ravdin, P. M., Siminoff, L. A., Davis, G. J., Mercer, M. B., Hewlett, J., Gerson, N., and Parker, H. L. (2001). Computer Program to Assist in Making Decisions About Adjuvant Therapy for Women With Early Breast Cancer. *Journal of Clinical Oncology*, 19(4):980–991. 679 citations (Crossref) [2021-08-06].

Rosenman, E., Basse, G., Owen, A., and Baiocchi, M. (2020). Combining Observational and Experimental Datasets Using Shrinkage Estimators. *arXiv:2002.06708 [math, stat]*. arXiv: 2002.06708.

Sterne, J. A. and Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology*, 54(10):1046–1055.

van Geloven, N., Swanson, S. A., Ramspek, C. L., Luijken, K., van Diepen, M., Morris, T. P., Groenwold, R. H. H., van Houwelingen, H. C., Putter, H., and le Cessie, S. (2020). Prediction meets causal inference: the role of treatment in clinical prediction models. *European Journal of Epidemiology*, 35(7):619–630.

Watson, T. (2007). Practitioner's Guide to Generalized Linear Models. Technical report, Towers Watson.

Xu, Z., Arnold, M., Stevens, D., Kaptoge, S., Pennells, L., Sweeting, M. J., Barrett, J., Di Angelantonio, E., and Wood, A. M. (2021). Prediction of Cardiovascular Disease Risk Accounting for Future Initiation of Statin Treatment. *American Journal of Epidemiology*, page kwab031.

A Appendix

A.1 Identification

We now prove that the assumption expressed in Equation 1 is not sufficient for identifying the interventional distribution $p(y=1|\mathrm{do}(t_x={t_x}'), \boldsymbol{x}=\boldsymbol{x'})$ from observational data using a simple example where all variables are binary and $\beta_0(\boldsymbol{x'})=\beta_0^*$ for all values of $\boldsymbol{x}$ meaning that the untreated risk does not vary with $\boldsymbol{x}$. In this setting, estimating an offset model amounts to estimating the log odds of the untreated risk β_0. We first derive an expression for the expected log likelihood as a function of β_0 $L(\beta_0)=E_{p_{\mathrm{obs}}(y,t_x,u)}[l(y,\hat{\pi}(t_x,u,\beta_0)]$ under the observational distribution in this example Then we show that the ground truth solution β_0^* is not a stationary point, proving our claim. Writing

$$
\begin{aligned}
p_u =& p(u=1)\\
p_{{t_x}'u'} =& p(t_x={t_x}', u=u') = p(t_x={t_x}'|u=u')p(u=u')\\
\pi_{{t_x}'u'} =& p(y=1|t_x={t_x}', u=u')
\end{aligned}
$$

Then the data generating mechanism is:

$$u, t_x \sim \mathcal{B}(p_{{t_x}'u'}), y \sim \mathcal{B}(\pi_{t_x u})$$

The ground truth solutions β_0^* and $\beta_{t_x}^*$ are:

$$p(y=1|\mathrm{do}(t_x=0)) = (1-p_u)\pi_{00} + p_u\pi_{01} = \sigma(\beta_0^*) \tag{11}$$

$$p(y=1|\mathrm{do}(t_x=1)) = (1-p_u)\pi_{10} + p_u\pi_{11} = \sigma(\beta_0^* + \beta_{t_x}^*) \tag{12}$$

The Bernoulli log-likelihood is

$$l(y|t_x,\beta_0,\beta_{t_x}) = y\log\sigma(\beta_0+\beta_{t_x}t_x) + (1-y)\log(1-\sigma(\beta_0+\beta_{t_x}t_x))$$

In offset models $\beta_{t_x}^*$ is assumed given a priori and β_0 is the only parameter, resulting in the following expression for $L(\beta_0)$:

$$
\begin{aligned}
L(\beta_0) =& p_{00}\left[\pi_{00}\log\sigma(\beta_0) + (1-\pi_{00})\log(1-\sigma(\beta_0))\right]\\
&+ p_{01}\left[\pi_{01}\log\sigma(\beta_0) + (1-\pi_{01})\log(1-\sigma(\beta_0))\right]\\
&+ p_{10}\left[\pi_{10}\log\sigma(\beta_0+\beta_{t_x}^*) + (1-\pi_{10})\log(1-\sigma(\beta_0+\beta_{t_x}^*))\right]\\
&+ p_{11}\left[\pi_{11}\log\sigma(\beta_0+\beta_{t_x}^*) + (1-\pi_{11})\log(1-\sigma(\beta_0+\beta_{t_x}^*))\right]
\end{aligned}
$$

Taking the derivative with respect to β_0, noting that $(\log\sigma(x))' = 1-\sigma(x)$, we get:

$$
\begin{aligned}
\frac{\partial L}{\partial \beta_0} =& p_{00}\left[\pi_{00}(1-\sigma(\beta_0)) - (1-\pi_{00})\sigma(\beta_0)\right]\\
&+ p_{01}\left[\pi_{01}(1-\sigma(\beta_0)) - (1-\pi_{01})\sigma(\beta_0)\right]\\
&+ p_{10}\left[\pi_{10}(1-\sigma(\beta_0+\beta_{t_x}^*)) - (1-\pi_{10})\sigma(\beta_0+\beta_{t_x}^*)\right]\\
&+ p_{11}\left[\pi_{11}(1-\sigma(\beta_0+\beta_{t_x}^*)) - (1-\pi_{11})\sigma(\beta_0+\beta_{t_x}^*)\right]
\end{aligned} \tag{13}
$$

We now plug in the ground truth solutions for $\beta_0^*, \beta_{t_x}^*$.

$$
\begin{aligned}
\frac{\partial L}{\partial \beta_0}(\beta_0=\beta_0^*) =& p_{00}\left[\pi_{00}(1-p_u\pi_{01}-(1-p_u)\pi_{00}) - (1-\pi_{00})(p_u\pi_{01}+(1-p_u)\pi_{00})\right]\\
&+ p_{01}\left[\pi_{01}(1-p_u\pi_{01}-(1-p_u)\pi_{00}) - (1-\pi_{01})(p_u\pi_{01}+(1-p_u)\pi_{00})\right]\\
&+ p_{10}\left[\pi_{10}(1-p_u\pi_{11}-(1-p_u)\pi_{10}) - (1-\pi_{10})(p_u\pi_{11}+(1-p_u)\pi_{10})\right]\\
&+ p_{11}\left[\pi_{11}(1-p_u\pi_{11}-(1-p_u)\pi_{10}) - (1-\pi_{11})(p_u\pi_{11}+(1-p_u)\pi_{10})\right]
\end{aligned}
$$

Removing terms that cancel out in each line results in

$$\begin{aligned}
=&p_{00}\left[p_u(\pi_{00}-\pi_{01})\right]\\
+&p_{01}\left[(1-p_u)(\pi_{01}-\pi_{00})\right]\\
+&p_{10}\left[p_u(\pi_{10}-\pi_{11})\right]\\
+&p_{11}\left[(1-p_u)(\pi_{11}-\pi_{10})\right]
\end{aligned}$$

Substituting back $p_{t_x{}'u'} = p(t_x = t_x{}'|u = u')p(u = u')$:

$$\begin{aligned}
=&p(t_x=0|u=0)(1-p_u)\left[p_u(\pi_{00}-\pi_{01})\right]\\
+&p(t_x=0|u=1)p_u\left[(1-p_u)(\pi_{01}-\pi_{00})\right]\\
+&p(t_x=1|u=0)(1-p_u)\left[p_u(\pi_{10}-\pi_{11})\right]\\
+&p(t_x=1|u=1)p_u\left[(1-p_u)(\pi_{11}-\pi_{10})\right]
\end{aligned}$$

Factoring out $p_u(1-p_u)$ and re-arranging we arrive at our result:

$$\begin{aligned}
\frac{\partial L}{\partial \beta_0}(\beta_0=\beta_0^*) = p_u(1-p_u)\big[&(\pi_{01}-\pi_{00})\left(p(t_x=0|u=1)-p(t_x=0|u=0)\right)+\\
&(\pi_{11}-\pi_{10})\left(p(t_x=1|u=1)-p(t_x=1|u=0)\right)\big]
\end{aligned}$$

If there is no confounding this expression is zero, but in general it is not which means that the ground truth solution β_0^* is not an optimum of the expected log-likelihood in the observational data distribution. This proves our claim that the offset model does not recover the interventional distribution in the presence of confounding. □

Of note, the fact that the interventional distribution is not identified does not automatically imply that the CATE is not identified as there may be another $\beta_0' \neq \beta_0^*$ such that $\text{CATE}(\beta_0 = \beta_0', \beta_{t_x} = \beta_{t_x}^*) = \text{CATE}(\beta_0 = \beta_0^*, \beta_{t_x} = \beta_{t_x}^*)$. To investigate this, assume that for some $\beta_0^* = a$ and $\beta_{t_x}^* = b$ we have that:

$$\begin{aligned}
\delta :=&\text{CATE}(\beta_0 = a, \beta_{t_x} = b)\\
=&\sigma(a+b)-\sigma(a)\\
=&\frac{e^{a+b}}{1+e^{a+b}}-\frac{e^a}{1+e^a}
\end{aligned}$$

Again, treating $\beta_{t_x}^*$ as fixed, we will now prove that this equation has at most two solutions for $\beta_0 = a$ by noting that:

$$\begin{aligned}
\frac{e^{a+b}}{1+e^{a+b}}-\frac{e^a}{1+e^a} &= \frac{e^{a+b}(1+e^a)-(1+e^{a+b})e^a}{(1+e^{a+b})(1+e^a)}\\
&= \frac{e^a(e^b-1)}{(1+e^{a+b})(1+e^a)}
\end{aligned}$$

Introducing $y := e^a$ and cross-multiplying we get:

$$\begin{aligned}
\delta = \frac{y(e^b-1)}{(1+e^b y)(1+y)} &\iff\\
\delta(1+e^b y)(1+y) = y(e^b-1) &=\\
\delta + \delta(1+e^b)y + \delta e^b y^2 = y(e^b-1) &\iff\\
\delta e^b y^2 + \left(\delta(1+e^b) - e^b + 1\right)y + \delta = 0&
\end{aligned}$$

Depending on the values of δ and b this quadratic equation in y has 0, 1 or 2 real-valued solutions, yielding 0, 1 or 2 real-valued solutions for $a = \log y = \beta_0$. This implies that there exists utmost one alternative solution $\beta_0' \neq \beta_0^*$ such that $\text{CATE}(\beta_0 = \beta_0', \beta_{t_x} = \beta_{t_x}^*) = \text{CATE}(\beta_0 = \beta_0^*, \beta_{t_x} = \beta_{t_x}^*)$.

In fact, we can explicitly compute this alternative solution by exploiting the symmetry of the sigmoid function: $\sigma(x) = 1 - \sigma(-x)$. Whenever it is true that:

$$\sigma(\beta_0^* + \beta_{t_x}^*) - \sigma(\beta_0^*) = \delta$$

It must simultaneously be true that, writing $\beta_0' := -(\beta_0^* + \beta_{t_x}^*)$:

$$\begin{aligned}
\sigma(\beta_0' + \beta_{t_x}^*) - \sigma(\beta_0') &= \\
\sigma(-(\beta_0^* + \beta_{t_x}^*) + \beta_{t_x}^*) - \sigma(-(\beta_0^* + \beta_{t_x}^*)) &= \\
\sigma(-\beta_0^*) - \sigma(-(\beta_0^* + \beta_{t_x}^*)) &= \\
(1 - \sigma(\beta_0^*)) - (1 - \sigma(\beta_0^* + \beta_{t_x}^*)) &= \\
\sigma(\beta_0^* + \beta_{t_x}^*) - \sigma(\beta_0^*) &= \delta
\end{aligned}$$

This means that except in the trivial case when $\beta_0^* = \beta_{t_x}^* = 0$ there *always* exists a second solution β_0' that has the same CATE δ but a different interventional distribution $p(y|\text{do}(t_x))$. We can check whether this coincidentally coincides with the maximum likelihood solution for β_0 in the offset model on the observational data by plugging in $\beta_0' := -(\beta^* + \beta_{t_x}^*)$ in the expression of the gradient of the likelihood (Equation 13). Again we remove terms that cancel out and substitute back $p_{t_x'u'} = p(t_x = t_x'|u = u')p(u = u')$ to arrive at:

$$\frac{\partial L}{\partial \beta_0}(\beta_0 = \beta_0') = p_u(1 - p_u)\big(p(t_x = 0|u = 1) - p(t_x = 0|u = 0)\big)\big((\pi_{10} - \pi_{11}) + (\pi_{01} - \pi_{00})\big) \tag{14}$$

$$+ p_u\big((\pi_{11} - \pi_{10}) + (\pi_{01} - \pi_{00})\big) \tag{15}$$

$$+ 2\pi_{10} + \pi_{11} - 1 \tag{16}$$

Analyzing this expression line-by-line we see that the first two lines are non-zero in general when there is confounding such that $p(t_x = 0|u = 1) \neq p(t_x = 0|u = 0)$ and $\pi_{t_x 1} \neq \pi_{t_x 0}$. The last line is also non-zero in general as $\pi_{t_x u}$ are free parameters.

A.2 Non-Collapsibility

Here we provide an example and intution on what non-collapsibility of the odds-ratio is and why it increases when the assocation between x and y becomes greater. Consider the following data-generating mechanism for binary x with $p(x = 1) = 0.5$, binary treatment t_x, and outcome mechanism $p(y = 1|\text{do}(t_x), x) = \sigma(\beta_0(x) + t_x)$, so that the *conditional* odds-ratio ($e^1 \approx 2.72$) is constant. As we will see, depending on how β_0 depends on x, the *marginal* log odds-ratio γ_{t_x} will vary. For two settings of $\beta_0(x)$ we calculate the resulting *marginal* odds-ratio γ_{t_x} in a few simple steps. The calculations are visualized in Figure 5. Let $\pi_{t_x}(x) = p(y = 1|\text{do}(t_x), x)$:

$$\begin{aligned}
\pi_0(0) &= \sigma(\beta_0(x = 0)) \\
\pi_0(1) &= \sigma(\beta_0(x = 1)) \\
\pi_1(0) &= \sigma(\beta_0(x = 0) + 1) \\
\pi_1(1) &= \sigma(\beta_0(x = 1) + 1) \\
\pi_0 &= (1 - p(x = 1))\pi_0(0) + p(x = 1)\pi_0(1) \\
\pi_1 &= (1 - p(x = 1))\pi_1(0) + p(x = 1)\pi_1(1) \\
\eta_0 &= \sigma^{-1}(\pi_0) \\
\eta_1 &= \sigma^{-1}(\pi_1) \\
\gamma_{t_x} &= \eta_1 - \eta_0
\end{aligned}$$

This leads to the following numerical results in Table 1 where we see that $\beta_{t_x} > \gamma_{t_x} > 0$ and $\gamma_{t_x} \to 0$ when the difference between $\pi_0(0), \pi_0(1)$ becomes bigger, despite $\beta_{t_x} = 1$ remaining constant.

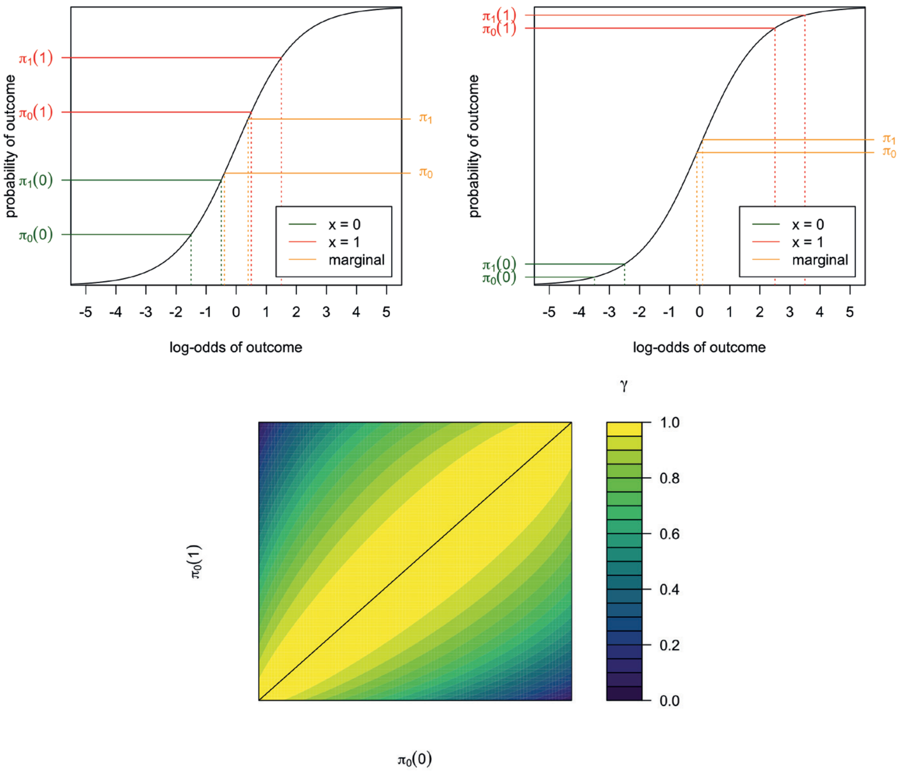


Figure 5: Illustration of non-collapsibility. For fixed $p(x = 1) = 0.5$ and $\beta_{t_x} = 1.0$, the marginal log odds-ratio γ_{t_x} of treatment becomes closer to 0 when the difference in the untreated risks $\pi_0(0), \pi_0(1)$ becomes larger.

setting	x	$\eta_0(x)$	$\eta_1(x)$	β_{t_x}	$\pi_0(x)$	$\pi_1(x)$	π_0	π_1	η_0	η_1	γ_{t_x}
a	0	-1.5	-0.5	1	0.182	0.378	0.402	0.598	-0.395	0.395	0.791
	1	0.5	1.5	1	0.622	0.818					
b	0	-3.5	-2.5	1	0.029	0.076	0.477	0.523	-0.093	0.093	0.186
	1	2.5	3.5	1	0.924	0.971					

Table 1

A.3 Additional Experiments

Extending the experiments in 3.2, we investigate the situation where x and u are correlated. Specifically, $p(x, u) = p(u|x)p(x)$ with $p(u|x = 0) = 1 - p(u|x = 1) = \alpha$ and $\alpha \in [0.1, 0.3, 0.5, 0.7, 0.9]$, $p(x) = 0.5$. As seen in Figure 6, whereas the PEHE for the marginal offset model increases with $\hat{\beta}_x$, the constrained offset model remains relatively unbiased in a wide range of settings. Again, the PEHE of the constrained offset model is always better than the fully observational baseline. In the areas with very high confounding ($\beta_{u\to y} \geq \log(5)$) and strong negative correlation between u and x ($\alpha \geq 0.7$) there are some settings where the constrained offset models perform worse than the ATE-baseline. As x and u both increase the probability of y but x and u are anti-correlated in these settings, we get close to the situation that $p(y|\text{do}(t_x), x = 0) \approx p(y|\text{do}(t_x), x = 1)$. This means that in these cases, the ATE-baseline has low PEHE as there is no actual difference in baseline risk depending on x. Whether this situation is relevant in actual applications will depend on the available background knowledge. The implication would be that the entire population under study would have the same outcome probability if they were included in a RCT and were assigned to the control arm. This total lack of variation in baseline risk may be deemed implausible in many concrete applications. Outside of these settings, the constrained offset models have better PEHE than the ATE-baseline whenever $\widehat{\text{OR}}_x \neq 1$.

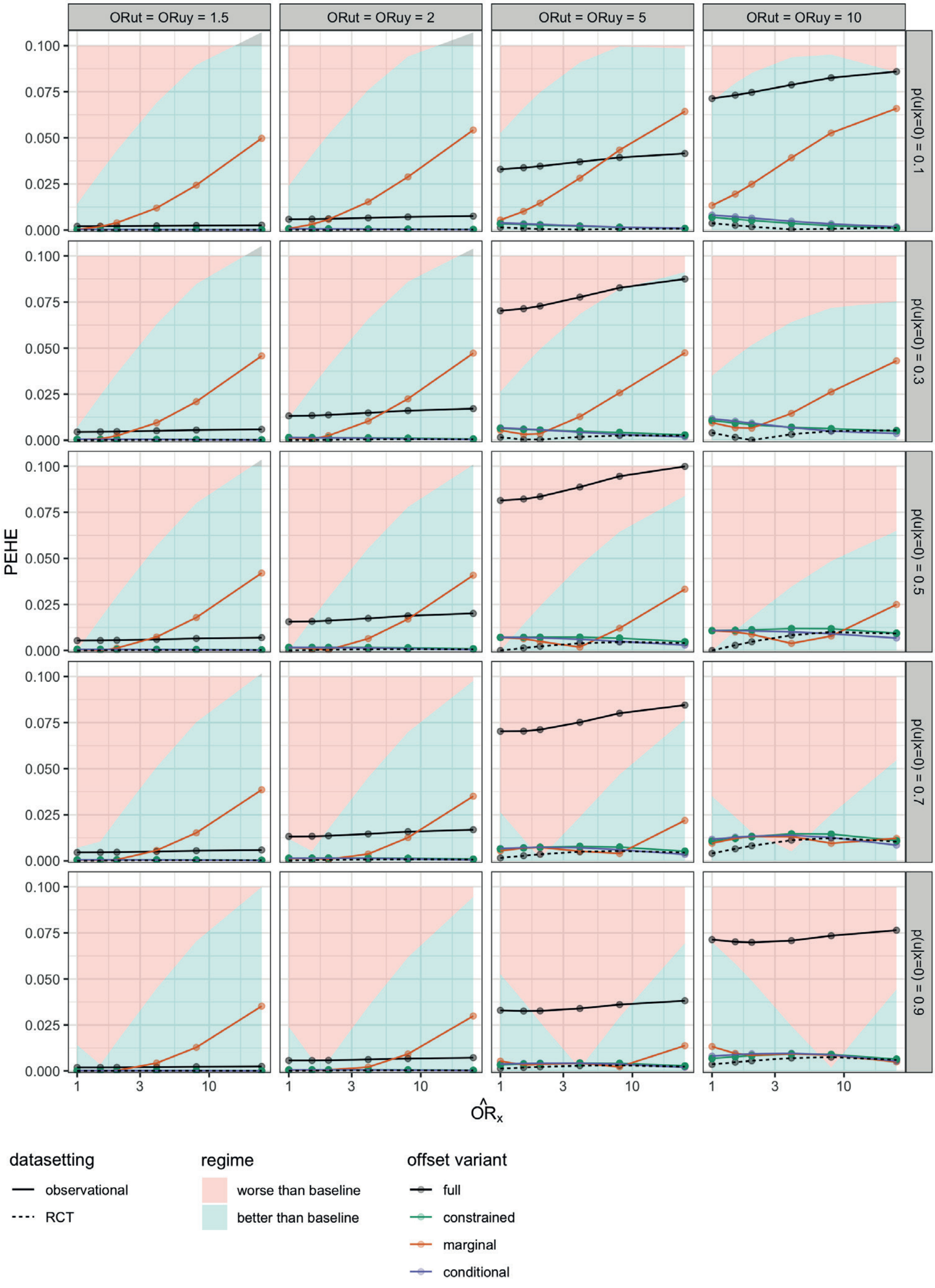


Figure 6: PEHEs for different offset models, indexed by $OR_{ut} = OR_{uy}$, the odds-ratios from confounder u to treatment t_x and outcome y respectively, and $p(u|x = 0) = 1 - p(u|x = 1)$. The fully observational baseline is sometimes not visible because the PEHE is higher than the maximum value on the y-axis.

CHAPTER 7

Decision making in cancer: causal questions require causal answers

Wouter A.C. van Amsterdam[1,2], Pim A. de Jong[1], Joost J.C. Verhoeff[1], Karijn Suijkerbuijk[1], Tim Leiner[3], Rajesh Ranganath[4]

[1] University Medical Center Utrecht, the Netherlands,
[2] Babylon Health, United Kingdom,
[3] Mayo Clinic, USA,
[4] New York University, USA

Key messages

- Much outcome prediction research is published with the aim to improve future treatment decisions, but fails to appreciate the causal nature of this task
- Because causal reasoning is ignored, decisions based on these prediction models can lead to substantial harm
- This holds even for prediction models that are found to be accurate in prospective validation studies
- To make outcome prediction research relevant to guiding treatment decisions, a better appreciation of causality is needed

The fundamental question guiding treatment decisions is "What is the chance of a good outcome, if we give treatment A or B, given that we know characteristics X about this patient". This clinical question targets the effect of giving treatment A or B and is therefore a *causal* question (1).

Oncologic treatment guidelines rely on evidence from randomised controlled trials (RCTs) which estimate the *average* treatment effect in a certain patient population. However, treatments are not equally effective in all patients thus knowing the *individual* treatment effect would empower better decisions. Hence, many studies in cancer aim to improve individual treatment decisions with outcome prediction models. These models use patient and tumor characteristics to predict a clinical outcome, such as overall survival. The motivation behind these models is that predictions from them can be useful for the task of selecting the best treatment for an individual patient. However, most outcome prediction research ignores the *causal* nature of selecting the best treatment and is thereby incapable of fulfilling its motivation.

We will demonstrate that non-causal outcome prediction models answer questions that render these models unsuited for decision-making. Outcome predictions can misguide treatment decisions, leading to worse patient outcomes, even when the predictions are accurate. Subsequently, we will provide directions on what is needed to estimate individual treatment effects, elaborating on the setting of non-randomized study designs.

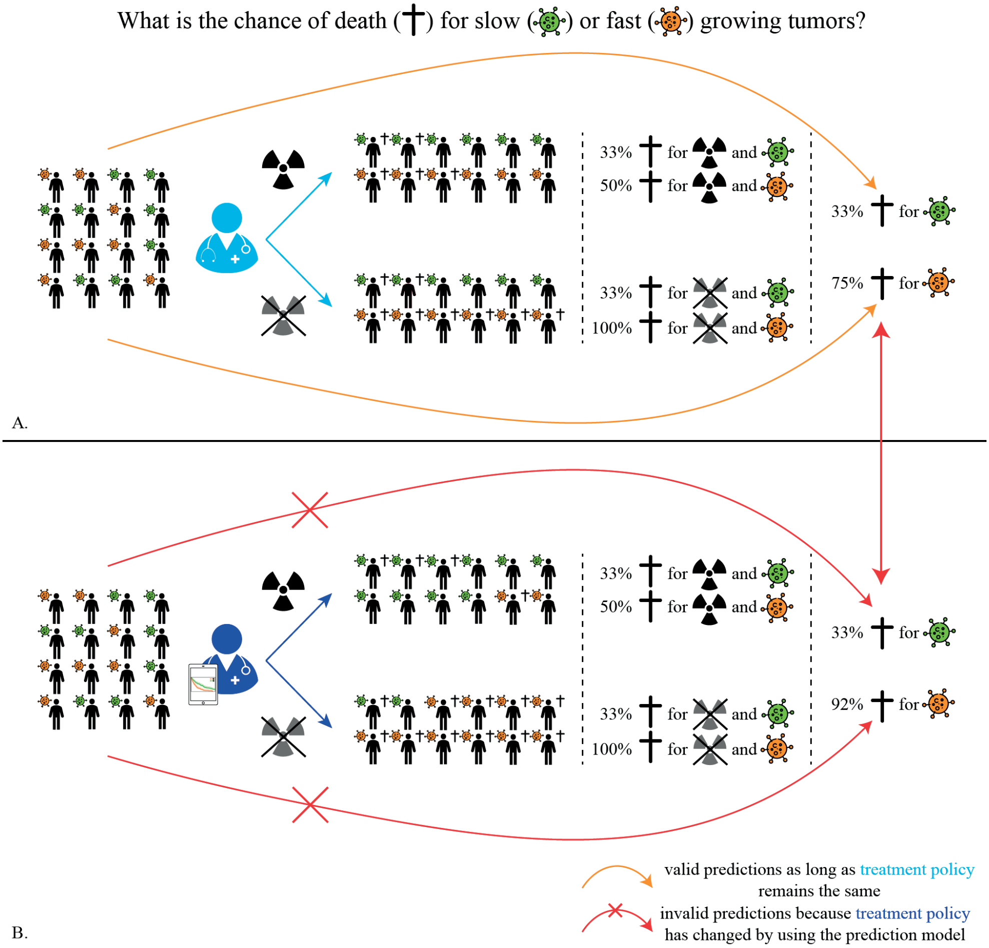
Non-causal prediction models are unsuited for decision support

To explain why non-causal outcome prediction models are unsuited for treatment decision support and how this introduces a substantial risk of harm due to misguided decisions, we first divide non-causal prediction models in three types: **treatment-naïve**, **post-decision** and **single-treatment**.

Treatment-naïve models (2–4) make use of baseline characteristics of a patient to predict the outcome. These models answer the question "What is the chance of a good outcome, given that we know X about this patient *with the assumption that we will keep making the same treatment decisions as we always did*". This assumption is necessary because if there would be a change in the way treatment decisions are made, for example by using the prediction model for more individualized treatment decisions, the patterns in the data are changed compared to when the model was developed and the predictions are no longer valid (Figure 1). These models may cause more harm than good when used to support treatment decisions.

As a simplified example consider a model that predicts overall survival for stage IV lung cancer patients based on the pretreatment growth rate of the tumor. Faster growing tumors generally lead to worse overall survival so an accurate model would predict a lower survival for patients with faster growing tumors. Applying this model, a clinician could decide to refrain from palliative radiotherapy in patients with faster growing tumors under the assumption that their life expectancy is too short to benefit from radiotherapy. This decision based on the non-causal prediction model would be unjustified and harmful, as faster growing tumors are more susceptible to radiotherapy (5). Thus, introducing this model in clinical practice as a decision support tool is likely to cause harm, even though the predictions are accurate.

Figure 1. Treatment-naïve models

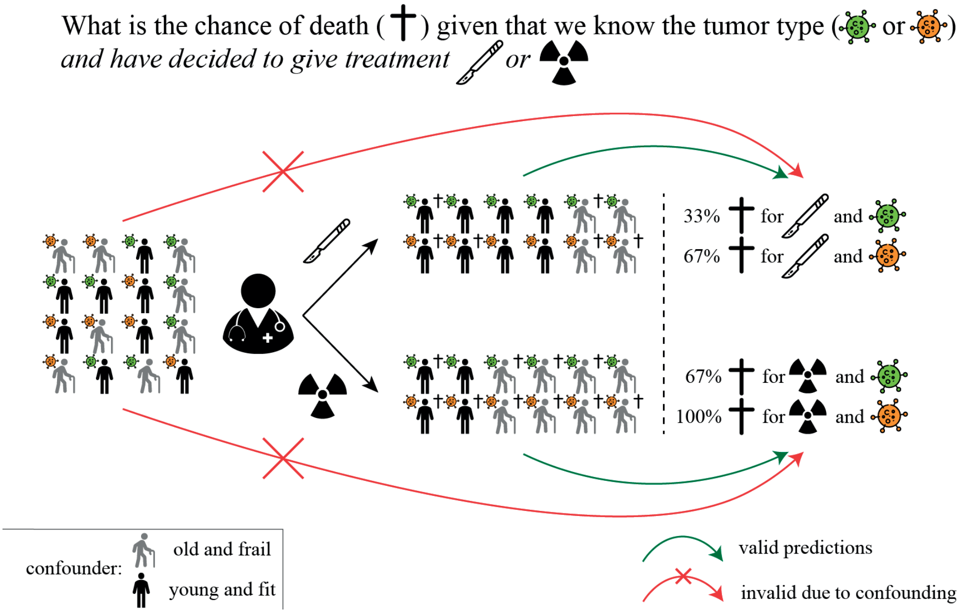

What is the chance of death (†) for slow or fast growing tumors?
33% † for and
50% † for and
33% † for and
100% † for and
33% † for
75% † for
A.
33% † for and
50% † for and
33% † for and
100% † for and
33% † for
92% † for
B.
valid predictions as long as treatment policy remains the same
invalid predictions because treatment policy has changed by using the prediction model

Treatment-naïve models predict the average outcome (death) depending on patient characteristics (growth rate) under the historic treatment policy (palliative radiotherapy or no radiotherapy). This means the predictions are only valid if the treatment policy does not change. In the first treatment policy (A.) under which the treatment-naïve model was developed, the decision for radiotherapy did not depend on tumor growth rate. In the second policy informed by the prediction model (B.), patients with faster growing tumors are less likely to get radiotherapy, even though radiotherapy would be particularly helpful for them. Despite being accurate, introducing the non-causal model for decision support has caused harm.

Post-decision models (6–8) incorporate historical treatments but do not estimate the causal effect of these treatments. Thereby they target the question: “What is the chance of a good outcome, given that we know X about this patient *with the assumption that the decision to give treatment A (or B) has already been made*”. In contrast with RCTs, treatment decisions in clinical practice are not made at random. This results in systematic differences between patient groups who underwent different treatments. When these systematic differences are not fully accounted for by the variables included in the prediction model and are at the same time related to the outcome, the problem of confounding occurs, sometimes referred to as confounding by indication or selection bias. Because of confounding, differences between outcome predictions from a post-decision model are not attributable to the causal effect of the treatment (Figure 2).

As a simple example, stage I non-small cell lung cancer (NSCLC) patients who are medically unfit for surgery due to advanced age and comorbidities will generally be treated with radiotherapy instead of surgery. Patients who underwent surgery for stage I NSCLC have better overall survival than patients who underwent radiotherapy, but this difference is not directly attributable to the *causal* effect of surgical treatment as the patient groups were not comparable to begin with. Importantly, the average survival of surgically treated patients is not the expected survival *if one were to operate on patients who are unfit for surgery*. Due to the confounding caused by the patient selection policy for surgery, the predictions of a post-decision model do not answer the question on what the outcome would be *if we would give treatment A or B*. Post-decision models should only be used when the treatment decision has already been made. Unaware of this caveat, a clinician or multi-disciplinary team may be tempted to use a post-decision model for a specific patient before the treatment decision has been made. If confounding makes the treatment appear more effective this could lead to overtreatment compared to the situation before introduction of the prediction model and vice versa. Many studies based on cancer registries such as the Surveillance, Epidemiology and End Results registry fall in the post-decision category.

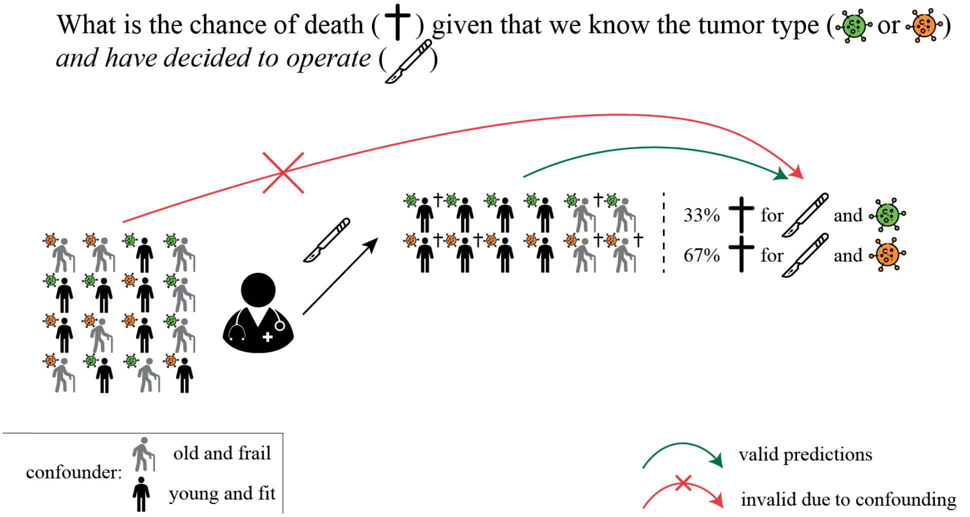
Figure 2. Post-decision models


What is the chance of death (†) given that we know the tumor type (or)
and have decided to give treatment *or*
33% † for and
67% † for and
67% † for and
100% † for and
confounder:
old and frail
young and fit
valid predictions
invalid due to confounding

Post-decision models predict the outcome (death) of patients depending on their characteristics (tumor type) in different treatment groups (surgery or radiotherapy) but are only valid after the treatment decision is made. This is because the distribution of the confounder (old and frail vs young and fit) differs between the two treated populations due to the treatment selection policy (young and fit patients are more likely to get surgery). As a result, neither treated populations are a valid reference group for the pretreatment population. Due to the differences in the distribution of the confounder, the model predictions are only valid in the populations in which the model was developed (the respective post-treatment decision populations), but not in the pretreatment population, making it unsuitable to guide treatment decisions.

Single-treatment models (9,10) are a special type of post-decision model that consider patients who had a single treatment of interest and answer the question: “What is the chance of a good outcome, given that we know X about this patient *and have decided to give treatment A*” (Figure 3). As there is no decision between multiple treatments here, it may seem that there can be no confounding. However, these models suffer from the same bias as multi-treatment post-decision models. Again, the population in which the model was developed, the patients who had treatment A, differs substantially from the population in which the treatment decision is not yet made with respect to the confounders. This makes the model unreliable in this wider population and likely to cause overtreatment or undertreatment. Many response prediction studies fall in the single-treatment category (9).

Figure 3. Single-treatment models


What is the chance of death ( ) given that we know the tumor type (or)
and have decided to operate ( )
33% for and
67% for and
confounder:
old and frail
young and fit
valid predictions
invalid due to confounding

Single-treatment models predict the outcome (death) under a certain treatment of interest (surgery) depending on the characteristics of a patient (tumor type). Just as post-decision models, single-treatment models are only valid in the population for whom the treatment decision has already been made, as the distribution of the confounder (old and frail vs young and fit) differs between the pre-treatment population and the population in which the treatment decision is already made due to the treatment policy (young and fit patients are more likely to get surgery).

Prospective validation is not sufficient

Prospective validation studies are the gold-standard for evaluating the accuracy of a prediction model (11), but they do not provide information on whether a prediction model is suitable for treatment decision support. In a prospective validation study, patient characteristics and outcomes are recorded for a new patient cohort according to a predefined protocol. Comparing the predictions of a model with the observed outcomes in these patients provides an estimate of how accurate the model is outside of the cohort in which the model was developed.

Introducing a prediction model to support treatment decisions constitutes an intervention that changes treatment decisions and thus patient outcomes. Prospective validation is not a suitable study design to test the effects of this intervention on treatment decisions and patient outcomes. In fact, even if a model is found to be accurate in a prospective validation study it could still lead to harm if it were used to guide treatment decisions. To illustrate this we continue with the earlier example of palliative radiotherapy for stage IV lung cancer where patients with fast growing tumors were recommended to not get radiotherapy. If this model were tested in a prospective validation study, the patients with fast-growing tumors would be given radiotherapy less often due to the treatment recommendation of the model, leading to even worse survival for these patients than before introduction of the prediction model. The introduction of the model has thus caused harm, but paradoxically it is still found to be accurate in the validation study as the model already predicted that patients with fast-growing tumors have a bad prognosis.

Hence, high accuracy in prospective validation studies is not sufficient evidence that a model is safe to use for guiding treatment decisions. Despite this, some guidelines on breast cancer (12) and prostate cancer (13) recommend usage of outcome prediction models for treatment decision support on the basis of prospective validation studies.

Developing individual treatment effect models in observational data

Now that we understand how non-causal prediction models fail to guide treatment decisions, the question remains: what research addresses this properly?

The ideal prediction model for supporting treatment decisions is an *individual treatment effect model.* Technically, a more appropriate term is the 'conditional average treatment effect', but as 'individual treatment effect' is more commonly used, we use this term instead. Individual treatment effect models estimate the effect of treatment for an individual patient based on their characteristics. Ideally, individual treatment effect models are estimated in data from RCTs (14).

However, the costs to have sufficient statistical power in all patient subgroups for all treatment protocols in RCTs are prohibitive. In addition, there are important settings in cancer care where treatment effect estimates from observational data are required. For example as provisional evidence to motivate a new trial; when specific patient subgroups are not represented in historic trials and when new biomarkers become available after the trials (15,16).

Individual treatment effect estimates from observational data require background knowledge and causal reasoning. The basic approach requires knowledge of the confounders (1). When some confounders are not available it is generally not possible to estimate the individual treatment effect, though there are important exceptions to this rule.

One such exception is when *proxy* measurements of the confounders are available (17–19). For example, in many cancer settings, the overall fitness of a patient is an important confounder (19). While there is no record of the overall fitness of a patient as assessed by the treating clinician, the performance score is a frequently available *proxy* measurement of the confounder overall fitness. In cases like these, dedicated proxy methods allow treatment effect estimation despite the unobserved confounder. Another situation is when instrumental variables are available (20,21). An instrumental variable is related to the treatment but is not confounded with the outcome, and influences the outcome only through the treatment. As an example, the general preference of a physician for a certain treatment strategy is sometimes used as an instrumental variable (22). Finally, methods that combine observational data and experimental data to further increase the efficiency of subgroup treatment effect estimates are under development and may yield interesting applications in cancer research (23,24).

Whatever approach is used, sensitivity analyses are an important tool for estimating the effects of potential violations of the made assumptions on the individual treatment effect estimates. For example, the effect of a potentially omitted confounder on the resulting treatment effect estimate can be calculated, resulting in a range of plausible treatment effects (25). Promising individual treatment effect models could be tested in *cluster-randomized trials*. In a cluster-randomized trial some groups of clinicians are randomly selected to get access to the model while others are not. In contrast with non-randomized validation studies, this allows for the estimation of the effect of introducing the model on treatment decisions and patient outcomes. Finally, the individual treatment effect is informative for the shared decision making process but needs to be put in perspective based on a patient's values and preferences.

Discussion

We highlighted that improving treatment decisions with prediction models is a fundamentally causal endeavour. Without causal reasoning, outcome prediction models cannot achieve the goal of improving treatment decisions. Based on unaddressed causal issues, non-causal prediction models that are found to be accurate can lead to worse decisions and patient harm. As the issue lies in causality, these problems cannot be resolved by larger datasets, more sophisticated prediction algorithms (e.g. machine learning) or even by prospective validation of prediction models.

In future research that aims to contribute to treatment decision support there should be more emphasis on addressing causal issues such as confounding instead of merely focusing on measures of predictive accuracy. Answering the fundamental question for shared treatment decision making "what is the chance of a good outcome if we give treatment A or B, given that we know X about this patient" cannot be done with ignorance of causality. Given the importance of this question, researchers, patient associations, journal editors, research funders and research consumers should prioritize research that addresses this over the multitude of non-causal prediction research.

References

[1] Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect. 2018.

[2] McMillan DC, Crozier JEM, Canna K, Angerson WJ, McArdle CS. Evaluation of an inflammation-based prognostic score (GPS) in patients undergoing resection for colon and rectal cancer. Int J Colorectal Dis. 2007 Jan 24;22(8):881.

[3] Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res. 2018 Mar 15;24(6):1248–59.

[4] Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, et al. Gene Expression Signature to Improve Prognosis Prediction of Stage II and III Colorectal Cancer. JCO. 2011 Jan 1;29(1):17–24.

[5] Breur K. Growth rate and radiosensitivity of human tumours—II: Radiosensitivity of human tumours. European Journal of Cancer (1965). 1966 Jun 1;2(2):173–88.

[6] Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer. 1997;79(4):857–62.

[7] Ryu J-S, Ryu HJ, Lee S-N, Memon A, Lee S-K, Nam H-S, et al. Prognostic impact of minimal pleural effusion in non-small-cell lung cancer. J Clin Oncol. 2014 Mar 20;32(9):960–7.

[8] Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, et al. Stage III Non-Small Cell Lung Cancer: Prognostic Value of FDG PET Quantitative Imaging Features Combined with Clinical Prognostic Factors. Radiology. 2016 Jan;278(1):214–22.

[9] Cappuzzo F, Hirsch FR, Rossi E, Bartolini S, Ceresoli GL, Bemis L, et al. Epidermal Growth Factor Receptor Gene and Protein and Gefitinib Sensitivity in Non–Small-Cell Lung Cancer. JNCI: Journal of the National Cancer Institute. 2005 May 4;97(9):643–55.

[10] Tendulkar RD, Agrawal S, Gao T, Efstathiou JA, Pisansky TM, Michalski JM, et al. Contemporary Update of a Multi-Institutional Predictive Nomogram for Salvage Radiotherapy After Radical Prostatectomy. J Clin Oncol. 2016 Oct 20;34(30):3648–54.

[11] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015 Jan 6;162(1):W1.

[12] Gradishar WJ. NCCN Breast Cancer Guideline, Version 5.2021 [Internet]. 2021 [cited 2021 Feb 15]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/breast-2.pdf

[13] Schaeffer E. NCCN Prostate Cancer Guideline, Version 2.2022 [Internet]. 2021 [cited 2022 Jan 7]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/prostate.pdf

[14] Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med. 2020 Jan 7;172(1):35–45.

[15] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002 Jan;415(6871):530–6.

[16] Cronin M, Pho M, Dutta D, Stephans JC, Shak S, Kiefer MC, et al. Measurement of Gene Expression in Archival Paraffin-Embedded Tissues: Development and Performance of a 92-Gene Reverse Transcriptase-Polymerase Chain Reaction Assay. The American Journal of Pathology. 2004 Jan 1;164(1):35–42.

[17] Kuroki M, Pearl J. Measurement bias and effect restoration in causal inference. Biometrika. 2014 Jun 1;101(2):423–37.

[18] Miao W, Geng Z, Tchetgen ET. Identifying Causal Effects With Proxy Variables of an Unmeasured Confounder. arXiv:160908816 [stat] [Internet]. 2018 Jun 28 [cited 2021 Jun 4]; Available from: http://arxiv.org/abs/1609.08816

[19] van Amsterdam WAC, Verhoeff JJC, Harlianto NI, Bartholomeus GA, Puli AM, de Jong PA, et al. Individual treatment effect estimation in the presence of unobserved confounding using proxies: a cohort study in stage III non-small cell lung cancer. Sci Rep. 2022 Apr 7;12(1):5848.

[20] Wald A. The Fitting of Straight Lines if Both Variables are Subject to Error. The Annals of Mathematical Statistics. 1940 Sep;11(3):284–300.

[21] Darolles S, Fan Y, Florens JP, Renault E. Nonparametric instrumental regression. Econometrica. 2011;79(5):1541–65.

[22] Chen Y, Briesacher BA. Use of instrumental variable in prescription drug research with observational data: a systematic review. Journal of Clinical Epidemiology. 2011 Jun 1;64(6):687–700.

[23] Rosenman E, Basse G, Owen A, Baiocchi M. Combining Observational and Experimental Datasets Using Shrinkage Estimators. arXiv:200206708 [math, stat] [Internet]. 2020 May 18 [cited 2021 Sep 3]; Available from: http://arxiv.org/abs/2002.06708

[24] Ilse M, Forré P, Welling M, Mooij JM. Efficient Causal Inference from Combined Observational and Interventional Data through Causal Reductions. arXiv:210304786 [cs, stat] [Internet]. 2021 Mar 8 [cited 2021 Sep 3]; Available from: http://arxiv.org/abs/2103.04786

[25] GREENLAND S. Basic Methods for Sensitivity Analysis of Biases. International Journal of Epidemiology. 1996 Dec 1;25(6):1107–16.

CHAPTER 8

Discussion

Abstract

Outcome prediction is popular in cancer research. The assumption motivating most of this research is that good outcome predictions help make better treatment decisions. In this discussion we emphasize that supporting treatment decisions is a causal task and thus requires causal reasoning. Most outcome prediction research in cancer ignores causal reasoning. This introduces a substantial risk of harm as decisions based on non-causal predictions can lead to worse patient outcomes, even if the predictions were found to be accurate in prospective validation studies. We illustrate potential causal issues using clinical examples. After highlighting these issues we explain what is needed to develop prediction models that can be useful for supporting treatment decisions. To make outcome prediction research relevant to supporting treatment decisions, a better appreciation of causality is needed.

Treatment decisions are guided by individual treatment effect estimates

Cancer is a disease with great variability in clinical outcomes, even within a single cancer type and cancer stage. Consequently, a large field of research is dedicated to predicting clinical outcomes for cancer patients. With the increasing popularity of machine learning methods and the availability of larger datasets, the interest in outcome prediction has grown even faster. In 1990, 3.4% of the PubMed matches for 'cancer' were also a match for 'prediction'. This fraction went up to 5.7% in 2000, steadily rising to 12.2% in 2021. The general assumption driving outcome prediction research in cancer is that accurate outcome predictions will allow patients and physicians to make better, more informed treatment decisions, ultimately improving patient outcomes.

Prediction models can support treatment decisions when they predict relevant clinical outcomes under different potential treatments. To date, few prediction models are recommended by oncologic treatment guidelines to support treatment decisions. To pass the scrutiny of treatment guideline requirements, the utility of a prediction model for treatment decisions needs to be clearly demonstrated, preferably in randomized controlled trials [1–3]. Notable examples are Oncoprint DX [4] and MammaPrint [5] for the selection of adjuvant therapy after surgical resection of breast cancer, and several molecular markers that are used to guide targeted therapy and immunotherapy in non-small cell lung cancer [1]. The fundamental principle of these models is the presence of treatment effect heterogeneity, meaning that the effect of treatment is not the same for all patients. We will use the term 'individual treatment effect model'[*] to denote prediction models that estimate the probability of an outcome under different hypothetical treatments based on pre-treatment patient characteristics. The fundamental question driving treatment decisions is: *"What is the probability of outcome Y if we would give treatment A (or B), given that we know X about this patient"*. This is a causal question and requires causal methods to answer.

Estimating the effect of treatments is not the target for many outcome prediction models [6–17]. A recent systematic review on prognostic factors for stage III non-small cell lung cancer **(chapter 2)** identified 65 studies. We re-examined the included studies to find whether the motivation of the prediction models was to improve future decisions and whether the studies appreciated the potential causal issues with how their research may address that aim. Out of 55 studies available through PubMed, 33 (60%) explicitly mentioned improving future treatment decisions as a motivation for the study [18–50]. Though all these studies were conducted in observational data, only 8 out of 33 (24%) noted that confounding may limit the validity of their results [19,24,27,28,34,35,45,50]. As we will explain later, confounding can render non-causal prediction models useless or even harmful for treatment decisions. The proportions are visualized in Figure 1. Though this review was limited to stage III non-small cell lung cancer, we have no reason to assume that these numbers will differ substantially in other cancer settings.

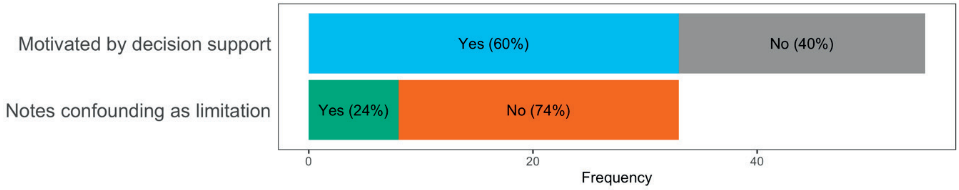


Figure 1. Frequency of motivation for prediction research being "to support treatment decisions" versus considerations of the potential issue of confounding with targeting this causal task. Studies are taken from ***chapter 2***.

In this discussion, we first point out the causal dimension of the assumption that good outcome predictions are useful for treatment decisions. We show that without appropriate appreciation of this causal dimension, accurate predictions can paradoxically lead to worse treatment decisions. To facilitate the discussion, we introduce three types of non-causal prediction models: **treatment-naïve**, **post-decision** and **single-treatment**. For each type, we describe what problems arise when using these models for decision making. We then describe what is needed to develop individual treatment effect models that have the potential to support future decisions.

---

[*] A technically more correct term is 'conditional average treatment effect model', but as individual treatment effect model is more used, we will use that term here as well.

Non-causal prediction models are unfit for decision making

Treatment-naïve models [6,8–13,18–37] make use of characteristics of a patient to predict the outcome. These models answer the question "What is the chance of a good outcome, given that we know X about this patient *with the assumption that we will keep making the same treatment decisions as we always did*". This assumption is necessary because if there would be a change in the way treatment decisions are made, for example by using the prediction model for more individualized treatment decisions, the patterns in the data are changed compared to when the model was developed and the predictions are no longer valid (Figure 2). These models may cause more harm than good when used to support treatment decisions.

As a simplified example consider a model that predicts overall survival for stage IV lung cancer patients based on the pretreatment growth rate of the tumor. Faster growing tumors generally lead to worse overall survival so an accurate model would predict a lower survival for patients with faster growing tumors. Applying this model, a clinician could decide to refrain from palliative radiotherapy in patients with faster growing tumors under the assumption that their life expectancy is too short to benefit from radiotherapy. This decision based on the non-causal prediction model would be unjustified and harmful, as faster growing tumors are more susceptible to radiotherapy [53]. Thus, introducing this model in clinical practice as a decision support tool is likely to cause harm, even though the predictions are accurate.

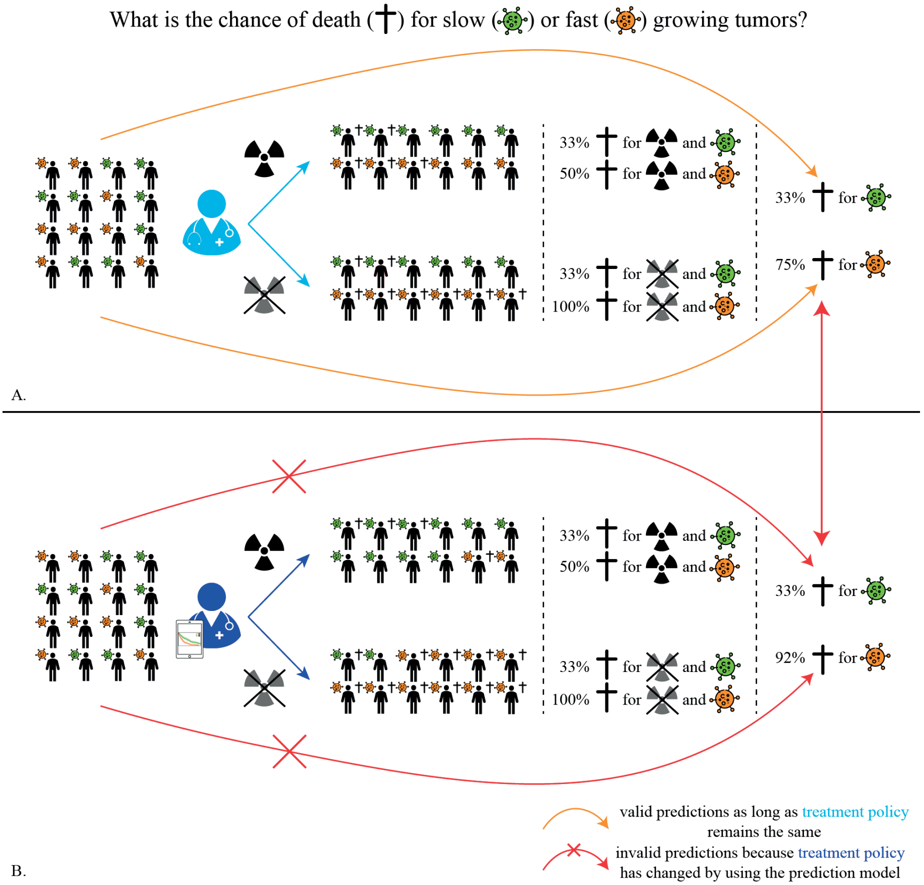

What is the chance of death (†) for slow (🟩) or fast (🟧) growing tumors?
33% † for ☢ and 🟩
50% † for ☢ and 🟧
33% † for 🟩
75% † for 🟧
33% † for ☒ and 🟩
100% † for ☒ and 🟧
A.
33% † for ☢ and 🟩
50% † for ☢ and 🟧
33% † for 🟩
92% † for 🟧
33% † for ☒ and 🟩
100% † for ☒ and 🟧
valid predictions as long as treatment policy remains the same
invalid predictions because treatment policy has changed by using the prediction model
B.

Figure 2. Treatment-naïve models predict the average outcome (death) depending on patient characteristics (growth rate) under the historic treatment policy (palliative radiotherapy or no radiotherapy). This means the predictions are only valid if the treatment policy does not change. In the first treatment policy (A.) under which the treatment-naïve model was developed, the decision for radiotherapy did not depend on tumor growth rate. In the second policy informed by the prediction model (B.), patients with faster growing tumors are less likely to get radiotherapy, even though radiotherapy would be particularly helpful for them. Despite being accurate, introducing the non-causal model for decision support has caused harm.

Post-decision models [7,38–45] incorporate historical treatments but do not estimate the causal effect of these treatments. Thereby they target the question: "What is the chance of a good outcome, given that we know X about this patient *with the assumption that the decision to give treatment A (or B) has already been made*". In contrast with RCTs, treatment decisions in clinical practice are not made at random. This results in systematic differences between patient groups who underwent different treatments. When these systematic differences are not fully accounted for by the variables included in the prediction model and are at the same time related to the outcome, the problem of confounding occurs, sometimes referred to as confounding by indication or selection bias. Because of confounding, differences between outcome predictions from a post-decision model are not attributable to the causal effect of the treatment (Figure 3).

As a simple example, stage I non-small cell lung cancer (NSCLC) patients who are medically unfit for surgery due to advanced age and comorbidities will generally be treated with radiotherapy instead of surgery. Patients who underwent surgery for stage I NSCLC have better overall survival than patients who underwent radiotherapy, but this difference is not directly attributable to the *causal* effect of surgical treatment as the patient groups were not comparable to begin with. Importantly, the average survival of surgically treated patients is not the expected survival *if one were to operate on patients who are unfit for surgery*. Due to the confounding caused by the patient selection policy for surgery, the predictions of a post-decision model do not answer the question on what the outcome would be *if we would give treatment A or B*. Post-decision models should only be used when the treatment decision has already been made. Unaware of this caveat, a clinician or multi-disciplinary team may be tempted to use a post-decision model for a specific patient before the treatment decision has been made. If confounding makes the treatment appear more effective this could lead to overtreatment compared to the situation before introduction of the prediction model and vice versa. Many studies based on cancer registries such as the Surveillance, Epidemiology and End Results registry fall in the post-decision category.

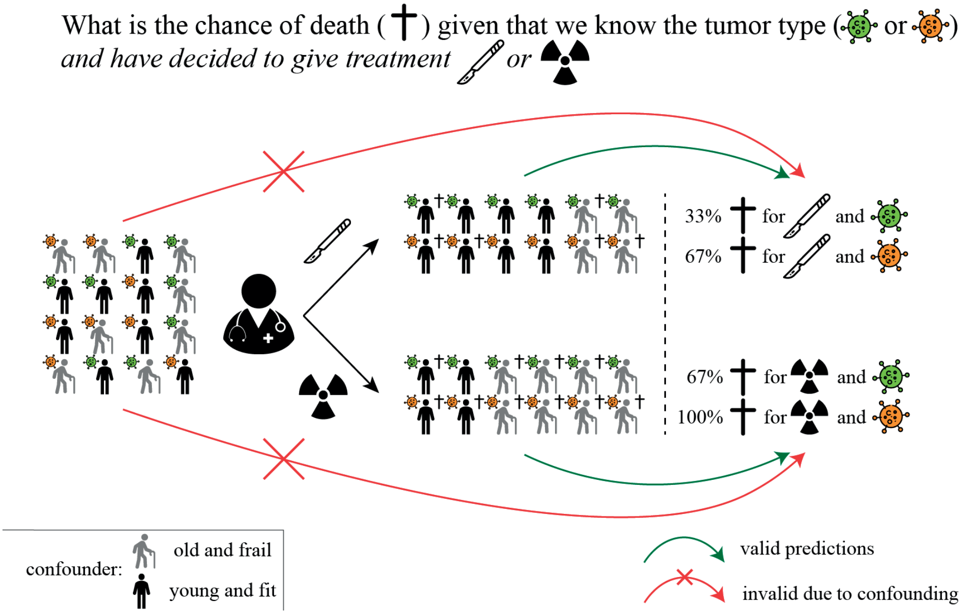

What is the chance of death (✝) given that we know the tumor type (or)
and have decided to give treatment or
33% ✝ for and
67% ✝ for and
67% ✝ for and
100% ✝ for and
confounder:
old and frail
young and fit
valid predictions
invalid due to confounding

Figure 3. Post-decision models predict the outcome (death) of patients depending on their characteristics (tumor type) in different treatment groups (surgery or radiotherapy) but are only valid after the treatment decision is made. This is because the distribution of the confounder (old and frail vs young and fit) differs between the two treated populations due to the treatment selection policy (young and fit patients are more likely to get surgery). As a result, neither treated populations are a valid reference group for the pretreatment population. Due to the differences in the distribution of the confounder, the model predictions are only valid in the populations in which the model was developed (the respective post-treatment decision populations), but not in the pretreatment population, making it unsuitable to guide treatment decisions.

Single-treatment models [54,55] are a special type of post-decision model that consider patients who had a single treatment of interest and answer the question: "What is the chance of a good outcome, given that we know X about this patient *and have decided to give treatment A*" (Figure 3). As there is no decision between multiple treatments here, it may seem that there can be no confounding. However, these models suffer from the same bias as multi-treatment post-decision models. Again, the population in which the model was developed, the patients who had treatment A, differs substantially from the population in which the treatment decision is not yet made with respect to the confounders. This makes the model unreliable in this wider population and likely to cause overtreatment or undertreatment. Many treatment response prediction studies fall in the single-treatment category [51,54].

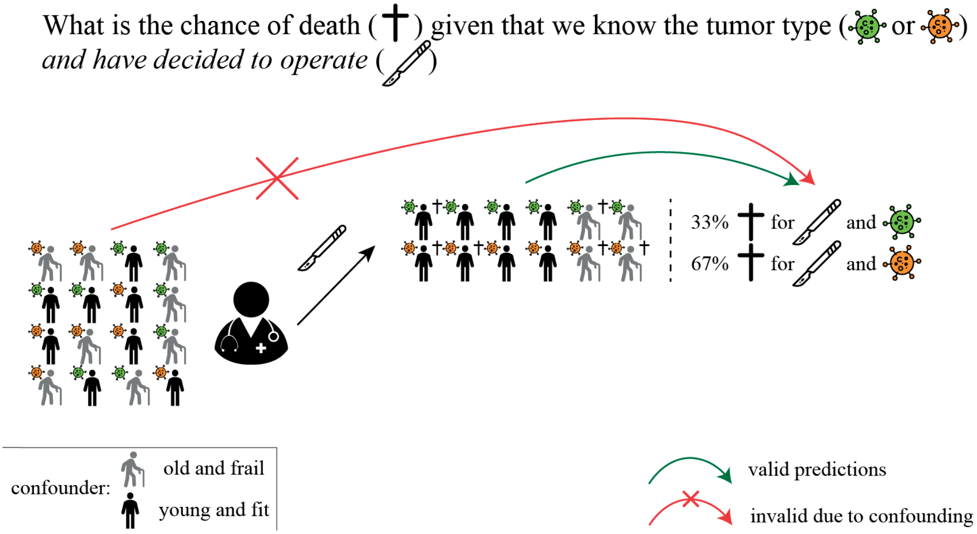

What is the chance of death (†) given that we know the tumor type (or)
and have decided to operate ( )
33% † for and
67% † for and
confounder:
old and frail
young and fit
valid predictions
invalid due to confounding

Figure 4. Single treatment models predict the outcome (death) under a certain treatment of interest (surgery) depending on the characteristics of a patient (tumor type). Just as post-decision models, single treatment models are only valid in the population for whom the treatment decision has already been made, as the distribution of the confounder (old and frail vs young and fit) differs between the pre-treatment population and the population in which the treatment decision is already made due to the treatment policy (young and fit patients are more likely to get surgery).

Prospective validation is not sufficient

Prospective validation studies are the gold-standard for evaluating the accuracy of a prediction model [56], but do not provide information on whether a prediction model is suitable for treatment decision support. In a prospective validation study, patient characteristics and outcomes are recorded for a new patient cohort according to a predefined protocol. Comparing the predictions of a model with the observed outcomes in these patients provides an estimate of how accurate the model is outside of the cohort in which the model was developed.

Introducing a prediction model to support treatment decisions constitutes an intervention that changes treatment decisions and thus patient outcomes. Prospective validation is not a suitable study design to test the effects of this intervention on treatment decisions and patient outcomes. In fact, even if a model is found to be accurate in a prospective validation study it could still lead to harm if it were used to guide treatment decisions. To illustrate this we continue with the earlier example of palliative radiotherapy for stage IV lung cancer where patients with fast growing tumors were recommended to not get radiotherapy. If this model were tested in a prospective validation study, the patients with fast-growing tumors would be given radiotherapy less often due to the treatment recommendation of the model, leading to even worse survival for these patients than before introduction of the prediction model. The introduction of the model has thus caused harm, but paradoxically it is still found to be accurate in the validation study as the model already predicted that patients with fast-growing tumors have a bad prognosis.

Hence, high accuracy in prospective validation studies is not sufficient evidence that a model is safe to use for guiding treatment decisions. Despite this, some guidelines on breast cancer [3] and prostate cancer [57] recommend usage of outcome prediction models for treatment decision support on the basis of prospective validation studies.

Developing individual treatment effect models in observational data

Now that we understand how non-causal prediction models cannot guide treatment decisions, the question remains: what research addresses this properly?

The ideal prediction model for supporting treatment decisions is an *individual treatment effect model.* Individual treatment effect models estimate the effect of treatment for an individual patient based on their characteristics. Ideally, individual treatment effect models are estimated in data from RCTs [58]. Guidelines for developing individual treatment effect models from randomized trial data are available in [58] and a tutorial is presented in [68]. However, the costs to have sufficient statistical power in all patient subgroups for all treatment protocols in RCTs are prohibitive. In addition, there are important settings in cancer care where treatment effect estimates from observational data are required. One such situation is when gathering pre-trial evidence on the efficacy of repurposed drugs for a new indication [59,60]. A second example is when treatment effect estimates are needed for patients who are not represented in the trials, for example older and weaker patients (**chapter 5**). As the trials provide no direct evidence on the effect of a treatment in these subpopulations, observational data may be used to find preliminary evidence of a treatment effect. Another example is when new biomarkers that are expected to be related to the treatment effect become available after the average treatment effect of a treatment has been established in randomized trials. The association between this new biomarker and the treatment effect must be studied in observational data [61,62] before there are enough grounds to study this association in new randomized trials [5,63].

Estimating treatment effects outside of RCTs (meaning from observational data) requires background knowledge and causal reasoning. The basic approach requires knowledge of the confounders [64]. When the confounders are known and accurately measured, there are many different approaches to estimating individual treatment effect models. These include outcome regression, inverse probability weighting methods, doubly robust estimators and other machine learning based estimators such as causal random forests [65] and neural network based estimators [66,67]. When some confounders are not available it is generally not possible to estimate the individual treatment effect, though there are important exceptions to this rule.

One such exception is when *proxy* measurements of the confounders are available (**chapter 5,** [69,70]). For example, in many cancer settings, the overall fitness of a patient is an important confounder (**chapter 5**). While there is generally no record of the overall fitness of a patient as assessed by the treating clinician, the *performance score* is a frequently available proxy measurement of the confounder overall fitness. In cases like these, dedicated proxy methods allow treatment effect estimation despite the unobserved confounder.

Another situation is when instrumental variables are available [71,72]. An instrumental variable is related to the treatment but is not confounded with the outcome, and influences the outcome only through the treatment. As an example, the general preference of a physician for a certain treatment strategy is sometimes used as an instrumental variable [73].

As a third exception, some individual treatment effect prediction models (*offset models*, **chapter 6**) use a treatment effect estimate from randomized trials on a relative scale (e.g. hazard ratio, risk ratio or odds ratio), combined with an untreated risk prediction model to estimate the individual treatment effect on an absolute risk scale. A treatment with a constant relative treatment effect has a different efficacy in different patients on an absolute risk scale based on the untreated risk of the patient. For example, suppose that randomized controlled trials have shown that heparin injections reduce the risk of thromboembolisms by a factor of 2. A patient with a 50% risk of developing thromboembolisms without prophylaxis will have a 25% reduction in absolute risk, whereas a patient with a 0.5% risk of developing thromboembolisms has a 0.25% reduction in absolute risk. Examples of offset models are Adjuvant! [74], Predict 2.0 [75] and Adjutorium [76], which predict the value of adjuvant therapy in breast cancer patients after surgical resection. We investigated offset models in **chapter 6** of this thesis and found that offset models are biased in the presence of confounding, though the resulting bias seems low. Based on statistical considerations we defined a refinement of offset models with a new constraint, and found that the constrained offset models have better performance than standard offset models.

Finally, methods that combine observational data and experimental data to further increase the efficiency of subgroup treatment effect estimates are under development and may yield interesting applications in cancer research [77,78].

Whatever approach is used, sensitivity analyses are an important tool for estimating the effects of violations of the made assumptions on the individual treatment effect estimates. For example, the effect of a potentially omitted confounder on the resulting treatment effect estimate can be calculated, resulting in a range of plausible treatment effects [79]. Promising individual treatment effect models could be tested in *cluster-randomized trials.* In a cluster-randomized trial some groups of clinicians are randomly selected to get access to the model while others are not. In contrast with non-randomized validation studies, this allows for the estimation of the effect of introducing the model on treatment decisions and patient outcomes.

Table 1. Overview of different prediction models.

Model type	**Question answered**	**Implied Assumption**	**Examples**
Treatment-naïve	"What is the probability of outcome Y given that we know X"	No change in treatment decision policy	(6,8–13,18–37)
Post-decision model	"What is the probability of outcome Y given that we know X and have decided to give treatment A or B"	Treatment decision already made	(7,38–45)
Single treatment post-decision model	"What is the probability of outcome Y given that we know X and have decided to give treatment A"	Treatment decision already made	(14–17,19,47–50)
Offset models	"What is the probability of outcome Y if we give treatment A or B given that we know X"	Constant relative treatment effect known from RCTs	(74,75), **chapter 6**
Individual treatment effect model	"What is the probability of outcome Y if we give treatment A or B, given that we know X"	Derived from RCT data or confounding otherwise addressed	(5,63), **chapter 5**

To highlight the differences between the different models in clinical practice, we include two hypothetical dialogues between a patient and an oncologist. The first is based on information from non-causal prediction models, the second is based on an individual treatment effect model derived from observational data.

Dialogue between patient and oncologist 1: non-causal models
Oncologist: Your work-up is done, we know your cancer type and stage

Patient: What is my prognosis?

Oncologist (treatment-naïve model): on average, other patients who share characteristics X with you live … more years.

Patient: Is there a treatment you can give me to improve my prognosis?

Oncologist: We know from randomized trials that treatment A leads to several more months survival than treatment B on average, though some patients do not respond well to treatment A and there may be severe side effects.

Patient: And how long do patients live with treatment A?

Oncologist: The average patient in the randomized trial who got treatment A lived … years, but those patients were younger and in better overall health than you so their results may not apply to your case.

Patient: So how long do patients like me survive when they get treatment A?

Oncologist (post-decision model): Looking back, patients who share characteristics X with you and got treatment A lived … years. However, I am not convinced that these patients are a good reference group for you as there are other characteristics Z that are important for survival but we have no information on Z for the historical patients.

Patient: This is getting a bit confusing, should I or should I not get treatment A?

Oncologist: I know this is a very tough decision, but ultimately it's yours to make.

Dialogue between patient and oncologist 2: individual treatment effect model
Oncologist: Your work-up is done, we know your cancer type and stage

Patient: What is my prognosis?

Oncologist (individual treatment effect model): That depends on what treatment strategy we will decide on. We cannot be exactly sure due to uncertainties associated with individual treatment effect estimation, but the best estimate based on our collective expertise and historical patient data is that you would on average live … years on treatment A, versus … years on treatment B.

Patient: Thank you for this information.

Conclusion

In this discussion we highlighted that improving treatment decisions with prediction models is a fundamentally causal endeavor. Much prediction research is published with the motivation to improve treatment decisions but causal the dimension of this task is often not acknowledged, let alone addressed. Without causal reasoning, prediction models cannot answer to the clinical motivation to improve treatment decisions. Based on unaddressed assumptions, non-causal prediction models that are found to be accurate in observational studies can lead to worse decisions and will likely fail to demonstrate value when evaluated in a randomized trial. As the issues lie in causality, they are not resolved by larger datasets, more flexible prediction algorithms (e.g. machine learning) or even by prospective validation of prediction models. In the ideal world from the eyes of individual treatment effect predictions there would be many large RCTs for all treatment comparisons where all important biomarkers are measured. In the reality, there are often only small RCTs in heterogeneous populations and with heterogeneity in treatment protocols. The trials are conducted without recording important biomarkers and are conducted in restricted subpopulations that may not reflect the entire target population. Therefore, treatment effect estimates from observational data remain needed in some settings.

Recommendations for future research

In outcome prediction research much weight is given to achieving high performance on metrics of prediction accuracy. As shown in this discussion, even prospectively validated highly accurate prediction models can cause more harm than good when used for decision making due to causal issues. In future research that aims to provide treatment decision support, there should be more emphasis on addressing causal issues such as confounding and treatment effect heterogeneity. A powerful way of formulating and presenting assumptions on confounders is with Directed Acyclic Graphs (DAG). In a DAG, the treatment, outcome and potential confounders are presented using nodes that depict variables, and arrows that depict causal dependencies between variables. A clear advantage of DAGs is that they provide a visual way of presenting the made assumptions to readers so that they may evaluate the appropriateness of these assumptions. If future individual treatment effect models are presented with DAGs, these DAGs can be discussed and expert consensus could arise on what DAGs are appropriate for what situations. A scaffold DAG that is amenable to many different cancer situations is presented in **chapter 5, Figure 1.** In addition to making the assumptions clear, performing sensitivity analyses as a standard practice will be highly beneficial as scientific claims regarding the utility of a new prediction model are contingent on the made assumptions. As unmeasured confounding such as caused by overall fitness will be a prevailing problem, proxy measurement methods and instrumental variable methods should be further developed and applied. If it is tenable to assume that a treatment has a constant effect on a relative scale, offset models are a very practical approach to estimating individual treatment effect models. Furthermore, prediction algorithms that use unstructured data such images directly, like neural networks, add an additional layer of complexity as it is no longer clear what information is in these images and how this information relates to known confounders (**chapter 4**)**.** Finally, methods that combine observational data and experimental data to further increase the efficiency of subgroup treatment effects are under development and may yield interesting applications in cancer research (77,78).

In conclusion, answering the fundamental question in treatment decisions "*what is the probability of outcome Y if we give treatment A or B, given that we know X about this patient*" is at the same time crucial for clinical practice and very challenging from a research perspective. Given the importance of this question, researchers, patient associations, journal editors, research funders and research consumers should prioritize research that addresses this over the multitude of non-causal prediction research.

References

[1] Ettinger DS. NCCN Non-Small Cell Lung Cancer Guideline, Version 1.2021 [Internet]. 2020 [cited 2021 Feb 15]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf

[2] Benson AB. NCCN Colon Cancer Guideline, Version 2.2021 [Internet]. 2021 [cited 2021 Feb 15]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/breast-2.pdf

[3] Gradishar WJ. NCCN Breast Cancer Guideline, Version 5.2021 [Internet]. 2021 [cited 2021 Feb 15]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/breast-2.pdf

[4] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004 Dec 30;351(27):2817–26.

[5] Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. New England Journal of Medicine. 2016 Aug 25;375(8):717–29.

[6] Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLOS Medicine. 2018 Nov 30;15(11):e1002711.

[7] Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer. 1997;79(4):857–62.

[8] Li X, Zhang Y, Zhang Y, Ding J, Wu K, Fan D. Survival prediction of gastric cancer by a seven-microRNA signature. Gut. 2010 May 1;59(5):579–85.

[9] McMillan DC, Crozier JEM, Canna K, Angerson WJ, McArdle CS. Evaluation of an inflammation-based prognostic score (GPS) in patients undergoing resection for colon and rectal cancer. Int J Colorectal Dis. 2007 Jan 24;22(8):881.

[10] Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res. 2018 Mar 15;24(6):1248–59.

[11] Proctor MJ, McMillan DC, Morrison DS, Fletcher CD, Horgan PG, Clarke SJ. A derived neutrophil to lymphocyte ratio predicts survival in patients with cancer. Br J Cancer. 2012 Aug;107(4):695–9.

[12] Ay C, Dunkler D, Marosi C, Chiriac A-L, Vormittag R, Simanek R, et al. Prediction of venous thromboembolism in cancer patients. Blood. 2010 Dec 9;116(24):5377–82.

[13] Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, et al. Gene Expression Signature to Improve Prognosis Prediction of Stage II and III Colorectal Cancer. JCO. 2011 Jan 1;29(1):17–24.

[14] Jensen GL, Yost CM, Mackin DS, Fried DV, Zhou S, Court LE, et al. Prognostic value of combining a quantitative image feature from positron emission tomography with clinical factors in oligometastatic non-small cell lung cancer. Radiotherapy and Oncology. 2018 Feb 1;126(2):362–7.

[15] Kattan MW, Potters L, Blasko JC, Beyer DC, Fearn P, Cavanagh W, et al. Pretreatment nomogram for predicting freedom from recurrence after permanent prostate brachytherapy in prostate cancer. Urology. 2001 Sep 1;58(3):393–9.

[16] Laurent-Puig P, Cayre A, Manceau G, Buc E, Bachet J-B, Lecomte T, et al. Analysis of PTEN, BRAF, and EGFR status in determining benefit from cetuximab therapy in wild-type KRAS metastatic colon cancer. Journal of Clinical Oncology. 2009;27(35):5924–30.

[17] Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. New England Journal of Medicine. 2002 Jun 20;346(25):1937–47.

18] Firat S, Byhardt RW, Gore E. Comorbidity and Karnofksy performance score are independent prognostic factors in stage III non-small-cell lung cancer: an institutional analysis of patients treated on four RTOG studies. Radiation Therapy Oncology Group. Int J Radiat Oncol Biol Phys. 2002 Oct 1;54(2):357–64.

19] Lee HY, Lee KS, Park J, Han J, Kim B-T, Kwon OJ, et al. Baseline SUVmax at PET-CT in stage IIIA non-small-cell lung cancer patients undergoing surgery after neoadjuvant therapy: prognostic implication focused on histopathologic subtypes. Acad Radiol. 2012 Apr;19(4):440–5.

[20] Gensheimer MF, Hong JC, Chang-Halpenny C, Zhu H, Eclov NCW, To J, et al. Mid-radiotherapy PET/CT for prognostication and detection of early progression in patients with stage III non-small cell lung cancer. Radiother Oncol. 2017 Nov;125(2):338–43.

[21] Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, et al. Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. Int J Radiat Oncol Biol Phys. 2014 Nov 15;90(4):834–42.

[22] Sibley GS, Mundt AJ, Shapiro C, Jacobs R, Chen G, Weichselbaum R, et al. The treatment of stage III nonsmall cell lung cancer using high dose conformal radiotherapy. Int J Radiat Oncol Biol Phys. 1995 Dec 1;33(5):1001–7.

[23] Soussan M, Chouahnia K, Maisonobe J-A, Boubaya M, Eder V, Morère J-F, et al. Prognostic implications of volume-based measurements on FDG PET/CT in stage III non-small-cell lung cancer after induction chemotherapy. Eur J Nucl Med Mol Imaging. 2013 May;40(5):668–76.

[24] Hyun SH, Ahn HK, Kim H, Ahn M-J, Park K, Ahn YC, et al. Volume-based assessment by (18)F-FDG PET/CT predicts survival in patients with stage III non-small-cell lung cancer. Eur J Nucl Med Mol Imaging. 2014 Jan;41(1):50–8.

[25] Wu J, Gensheimer MF, Dong X, Rubin DL, Napel S, Diehn M, et al. Robust Intratumor Partitioning to Identify High-Risk Subregions in Lung Cancer: A Pilot Study. Int J Radiat Oncol Biol Phys. 2016 Aug 1;95(5):1504–12.

[26] Elsayad K, Samhouri L, Scobioala S, Haverkamp U, Eich HT. Is tumor volume reduction during radiotherapy prognostic relevant in patients with stage III non-small cell lung cancer? J Cancer Res Clin Oncol. 2018 Jun;144(6):1165–71.

[27] Oberije C, De Ruysscher D, Houben R, van de Heuvel M, Uyterlinde W, Deasy JO, et al. A Validated Prediction Model for Overall Survival From Stage III Non-Small Cell Lung Cancer: Toward Survival Prediction for Individual Patients. Int J Radiat Oncol Biol Phys. 2015 Jul 15;92(4):935–44.

[28] Warner A, Dahele M, Hu B, Palma DA, Senan S, Oberije C, et al. Factors Associated With Early Mortality in Patients Treated With Concurrent Chemoradiation Therapy for Locally Advanced Non-Small Cell Lung Cancer. Int J Radiat Oncol Biol Phys. 2016 Mar 1;94(3):612–20.

[29] Hwang IG, Ahn MJ, Park BB, Ahn YC, Han J, Lee S, et al. ERCC1 expression as a prognostic marker in N2(+) nonsmall-cell lung cancer patients treated with platinum-based neoadjuvant concurrent chemoradiotherapy. Cancer. 2008 Sep 15;113(6):1379–86.

[30] Betticher DC, Hsu Schmitz S-F, Tötsch M, Hansen E, Joss C, von Briel C, et al. Prognostic factors affecting long-term outcomes in patients with resected stage IIIA pN2 non-small-cell lung cancer: 5-year follow-up of a phase II study. Br J Cancer. 2006 Apr 24;94(8):1099–106.

[31] Kanzaki H, Kataoka M, Nishikawa A, Uwatsu K, Nagasaki K, Nishijima N, et al. Impact of early tumor reduction on outcome differs by histological subtype in stage III non-small-cell lung cancer treated with definitive radiotherapy. Int J Clin Oncol. 2016 Oct;21(5):853–61.

[32] Casiraghi M, Guarize J, Sandri A, Maisonneuve P, Brambilla D, Romano R, et al. Pneumonectomy in Stage IIIA-N2 NSCLC: Should It Be Considered After Neoadjuvant Chemotherapy? Clin Lung Cancer. 2019 Mar;20(2):97-106.e1.

[33] Kim E, Wu H-G, Keam B, Kim TM, Kim D-W, Paeng JC, et al. Significance of (18)F-FDG PET Parameters According to Histologic Subtype in the Treatment Outcome of Stage III Non-small-cell Lung Cancer Undergoing Definitive Concurrent Chemoradiotherapy. Clin Lung Cancer. 2019 Jan;20(1):e9–23.

[34] Yoo GS, Oh D, Pyo H, Ahn YC, Noh JM, Park HC, et al. Concurrent chemo-radiotherapy for unresectable non-small cell lung cancer invading adjacent great vessels on radiologic findings: is it safe? J Radiat Res. 2019 Mar 1;60(2):234–41.

[35] Kim D-Y, Song C, Kim SH, Kim YJ, Lee JS, Kim J-S. Chemoradiotherapy versus radiotherapy alone following induction chemotherapy for elderly patients with stage III lung cancer. Radiat Oncol J. 2019 Sep;37(3):176–84.

[36] Hyun SH, Ahn HK, Ahn M-J, Ahn YC, Kim J, Shim YM, et al. Volume-Based Assessment With 18F-FDG PET/CT Improves Outcome Prediction for Patients With Stage IIIA-N2 Non-Small Cell Lung Cancer. AJR Am J Roentgenol. 2015 Sep;205(3):623–8.

[37] Morgensztern D, Waqar S, Subramanian J, Gao F, Trinkaus K, Govindan R. Prognostic significance of tumor size in patients with stage III non-small-cell lung cancer: a surveillance, epidemiology, and end results (SEER) survey from 1998 to 2003. J Thorac Oncol. 2012 Oct;7(10):1479–84.

[38] Ryu J-S, Ryu HJ, Lee S-N, Memon A, Lee S-K, Nam H-S, et al. Prognostic impact of minimal pleural effusion in non-small-cell lung cancer. J Clin Oncol. 2014 Mar 20;32(9):960–7.

[39] Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, et al. Stage III Non-Small Cell Lung Cancer: Prognostic Value of FDG PET Quantitative Imaging Features Combined with Clinical Prognostic Factors. Radiology. 2016 Jan;278(1):214–22.

[40] Alexander BM, Othus M, Caglar HB, Allen AM. Tumor volume is a prognostic factor in non-small-cell lung cancer treated with chemoradiotherapy. Int J Radiat Oncol Biol Phys. 2011 Apr 1;79(5):1381–7.

[41] Etiz D, Marks LB, Zhou S-M, Bentel GC, Clough R, Hernando ML, et al. Influence of tumor volume on survival in patients irradiated for non-small-cell lung cancer. Int J Radiat Oncol Biol Phys. 2002 Jul 15;53(4):835–46.

[42] Hayakawa K, Mitsuhashi N, Saito Y, Furuta M, Nakayama Y, Katano S, et al. Impact of tumor extent and location on treatment outcome in patients with stage III non-small cell lung cancer treated with radiation therapy. Jpn J Clin Oncol. 1996 Aug;26(4):221–8.

[43] Mao Q, Xia W, Dong G, Chen S, Wang A, Jin G, et al. A nomogram to predict the survival of stage IIIA-N2 non-small cell lung cancer after surgery. Vol. 155. United States; 2018.

[44] Dieleman EMT, Uitterhoeve ALJ, van Hoek MW, van Os RM, Wiersma J, Koolen MGJ, et al. Concurrent Daily Cisplatin and High-Dose Radiation Therapy in Patients With Stage III Non-Small Cell Lung Cancer. Int J Radiat Oncol Biol Phys. 2018 Nov 1;102(3):543–51.

[45] Tao X, Yuan C, Zheng D, Ye T, Pan Y, Zhang Y, et al. Outcomes comparison between neoadjuvant chemotherapy and adjuvant chemotherapy in stage IIIA non-small cell lung cancer patients. J Thorac Dis. 2019 Apr;11(4):1443–55.

[46] Lee VHF, Chan WWL, Lee EYP, Choy T-S, Ho PPY, Leung DKC, et al. Prognostic Significance of Standardized Uptake Value of Lymph Nodes on Survival for Stage III Non-small Cell Lung Cancer Treated With Definitive Concurrent Chemoradiotherapy. Am J Clin Oncol. 2016 Aug;39(4):355–62.

[47] Ahn SY, Park CM, Park SJ, Kim HJ, Song C, Lee SM, et al. Prognostic value of computed tomography texture features in non-small cell lung cancers treated with definitive concomitant chemoradiotherapy. Invest Radiol. 2015 Oct;50(10):719–25.

[48] Xiang Z-L, Erasmus J, Komaki R, Cox JD, Chang JY. FDG uptake correlates with recurrence and survival after treatment of unresectable stage III non-small cell lung cancer with high-dose proton therapy and chemotherapy. Radiat Oncol. 2012 Aug 28;7:144.

[49] Agrawal V, Coroller TP, Hou Y, Lee SW, Romano JL, Baldini EH, et al. Lymph node volume predicts survival but not nodal clearance in Stage IIIA-IIIB NSCLC. PLoS One. 2017;12(4):e0174268.

[50] Maniwa T, Shintani Y, Okami J, Kadota Y, Takeuchi Y, Takami K, et al. Upfront surgery in patients with clinical skip N2 lung cancer based on results of modern radiological examinations. J Thorac Dis. 2018 Dec;10(12):6828–37.

[51] Bensch F, van der Veen EL, Lub-de Hooge MN, Jorritsma-Smit A, Boellaard R, Kok IC, et al. 89Zr-atezolizumab imaging as a non-invasive approach to assess clinical response to PD-L1 blockade in cancer. Nat Med. 2018 Dec;24(12):1852–8.

[52] Trebeschi S, Drago SG, Birkbak NJ, Kurilova I, Călin AM, Delli Pizzi A, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. Ann Oncol. 2019 Jun 1;30(6):998–1004.

[53] Breur K. Growth rate and radiosensitivity of human tumours—II: Radiosensitivity of human tumours. European Journal of Cancer (1965). 1966 Jun 1;2(2):173–88.

[54] Cappuzzo F, Hirsch FR, Rossi E, Bartolini S, Ceresoli GL, Bemis L, et al. Epidermal Growth Factor Receptor Gene and Protein and Gefitinib Sensitivity in Non–Small-Cell Lung Cancer. JNCI: Journal of the National Cancer Institute. 2005 May 4;97(9):643–55.

[55] Tendulkar RD, Agrawal S, Gao T, Efstathiou JA, Pisansky TM, Michalski JM, et al. Contemporary Update of a Multi-Institutional Predictive Nomogram for Salvage Radiotherapy After Radical Prostatectomy. J Clin Oncol. 2016 Oct 20;34(30):3648–54.

[56] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015 Jan 6;162(1):W1.

[57] Schaeffer E. NCCN Prostate Cancer Guideline, Version 2.2022 [Internet]. 2021 [cited 2022 Jan 7]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/prostate.pdf

[58] Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med. 2020 Jan 7;172(1):35–45.

[59] Singhal S, Mehta J, Desikan R, Ayers D, Roberson P, Eddlemon P, et al. Antitumor activity of thalidomide in refractory multiple myeloma. New England Journal of Medicine. 1999;341(21):1565–71.

[60] Aggarwal S, Verma SS, Aggarwal S, Gupta SC. Drug repurposing for breast cancer therapy: Old weapon for new battle. Seminars in Cancer Biology. 2021 Jan 1;68:8–20.

[61] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002 Jan;415(6871):530–6.

[62] Cronin M, Pho M, Dutta D, Stephans JC, Shak S, Kiefer MC, et al. Measurement of Gene Expression in Archival Paraffin-Embedded Tissues: Development and Performance of a 92-Gene Reverse Transcriptase-Polymerase Chain Reaction Assay. The American Journal of Pathology. 2004 Jan 1;164(1):35–42.

[63] Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. New England Journal of Medicine. 2018 Jul 12;379(2):111–21.

[64] Pearl J. Causality. Cambridge University Press; 2009.

[65] Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association. 2018 Jul 3;113(523):1228–42.

[66] Shi C, Blei DM, Veitch V. Adapting Neural Networks for the Estimation of Treatment Effects. arXiv:190602120 [cs, stat] [Internet]. 2019 Oct 17 [cited 2021 Sep 3]; Available from: http://arxiv.org/abs/1906.02120

[67] Shalit U, Johansson FD, Sontag D. Estimating individual treatment effect: generalization bounds and algorithms. arXiv:160603976 [cs, stat] [Internet]. 2017 May 16 [cited 2021 Sep 23]; Available from: http://arxiv.org/abs/1606.03976

[68] Hoogland J, IntHout J, Belias M, Rovers MM, Riley RD, E Harrell F, et al. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. Stat Med. 2021 Aug 16;

[69] Kuroki M, Pearl J. Measurement bias and effect restoration in causal inference. Biometrika. 2014 Jun 1;101(2):423–37.

[70] Miao W, Geng Z, Tchetgen ET. Identifying Causal Effects With Proxy Variables of an Unmeasured Confounder. arXiv:160908816 [stat] [Internet]. 2018 Jun 28 [cited 2021 Jun 4]; Available from: http://arxiv.org/abs/1609.08816

[71] Wald A. The Fitting of Straight Lines if Both Variables are Subject to Error. The Annals of Mathematical Statistics. 1940 Sep;11(3):284–300.

[72] Darolles S, Fan Y, Florens JP, Renault E. Nonparametric instrumental regression. Econometrica. 2011;79(5):1541–65.

[73] Chen Y, Briesacher BA. Use of instrumental variable in prescription drug research with observational data: a systematic review. Journal of Clinical Epidemiology. 2011 Jun 1;64(6):687–700.

[74] Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, et al. Computer Program to Assist in Making Decisions About Adjuvant Therapy for Women With Early Breast Cancer. JCO. 2001 Feb 15;19(4):980–91.

[75] Candido dos Reis FJ, Wishart GC, Dicks EM, Greenberg D, Rashbass J, Schmidt MK, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. Breast Cancer Res. 2017 Dec;19(1):58.

[76] Alaa AM, Gurdasani D, Harris AL, Rashbass J, van der Schaar M. Machine learning to guide the use of adjuvant therapies for breast cancer. Nat Mach Intell. 2021 Jun 24;1–11.

[77] Rosenman E, Basse G, Owen A, Baiocchi M. Combining Observational and Experimental Datasets Using Shrinkage Estimators. arXiv:200206708 [math, stat] [Internet]. 2020 May 18 [cited 2021 Sep 3]; Available from: http://arxiv.org/abs/2002.06708

[78] Ilse M, Forré P, Welling M, Mooij JM. Efficient Causal Inference from Combined Observational and Interventional Data through Causal Reductions. arXiv:210304786 [cs, stat] [Internet]. 2021 Mar 8 [cited 2021 Sep 3]; Available from: http://arxiv.org/abs/2103.04786

[79] GREENLAND S. Basic Methods for Sensitivity Analysis of Biases. International Journal of Epidemiology. 1996 Dec 1;25(6):1107–16.

[80] van Amsterdam WAC, Verhoeff JJC, Harlianto NI, Bartholomeus GA, Puli AM, Jong PA de, et al. Treatment effect estimation in the presence of unobserved confounding using proxies: a study of non-small cell lung cancer. submitted.

[81] van Amsterdam WAC, Verhoeff JJC, de Jong PA, Leiner T, Eijkemans MJC. Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning. npj Digital Medicine. 2019 Dec 10;2(1):1–6.

CHAPTER 9

General Summary

General Summary

In this thesis we studied the prediction of overall survival in non-small cell lung cancer and the estimation of individual treatment effects. **Part 1** focusses on the prediction of overall survival. **Part 2** is dedicated to estimating individual treatment effects and includes studies on the methodological challenges of this task as well as a concrete application to stage III non-small cell lung cancer.

Part 1: predicting overall survival

In **chapter 2** we present a study that summarizes the literature on prognostic factors measurable on CT scans for stage III non-small cell lung cancer. With a systematic search strategy 65 articles were included, describing 26 unique prognostic factors and including 144,513 patients. There was a wide variation in study quality and only 4 studies compared the added value of potential new prognostic factors measured on CT scans with the standard available prognostic factors. This summary of the literature indicates that the total volume of the primary tumor and lymph node metastases, tumor diameter, volume of lymph node metastases and presence of pleural fluid are related to survival. Recommendations from this study are that future studies on prognostic factors should better report their method and results, and compare potential new prognostic factors with known prognostic factors.

In **chapter 3** we study the importance of muscle quantity and muscle density for overall survival in non-small cell lung cancer. A commonly used measure of the amount of muscle volume of a patient on CT scans is the cross-sectional muscle area of the psoas muscle on the third lumbar vertebra. This psoas muscle area correlates strongly with the total amount of muscle volume in the entire body. By dividing the psoas muscle area by the square of a patient's height, one gets the 'psoas muscle index' (PMI). A link between PMI and survival has been demonstrated in many cancers. Patients with a higher PMI (more muscle) live longer on average, presumably because they are in better general condition and/or because the tumor has taken a lesser toll on the body at the time of diagnosis. In addition to the amount of muscle, the quality of the muscle is also important. Through various processes, including aging, immobility and inflammatory reactions related to cancer, muscle tissue can be slowly replaced by fat tissue, making the muscle weaker. Fat tissue has a lower density than muscle tissue, the density of a muscle as measured on a CT scan is thus a measure of the amount of fat infiltration into that muscle and a proxy measure of muscle quality. The average density of the psoas muscle as seen on a CT scan at the level of the 3rd lumbar vertebra is a standardized measurement called psoas muscle radiodensity (PMD).

Several studies have examined the association between muscle area and muscle density and survival in lung cancer patients, but the studies have conflicting results. In **chapter 3** we present a possible explanation for the contradictions with a new hypothesis that the association between PMI and survival is stronger if muscle density, measured in PMD, is sufficient. The biological motivation behind this hypothesis is that more muscle volume (PMI) is only associated with better survival if the muscle quality, measured by proxy with PMD, is sufficient. In statistical terms, this means that there is a statistical interaction between PMI and PMD and survival. If the hypothesis is true, and if the patients in the previously published studies differ on average with respect to PMD, this could explain that the relationship found between PMI and survival differs per study. To answer this question, we collected a large group of 2480 non-small cell lung cancer patients who were treated at the radiotherapy department of the University Medical Center Utrecht. An automated computer algorithm based on a technique called 'deep learning' was used to measure the PMI and PMD in these patients. The association between PMI, PMD and survival was analyzed in the context of known tumor characteristics (histological subtype) and patient characteristics (age, sex, performance score and BMI). There was clear statistical evidence for our hypothesis. This means that future studies on the association between muscle quantity and overall survival should accommodate the effect that muscle radiodensity has on the association between muscle quantity and overall survival. Though our study was conducted in non-small cell lung cancer patients only, given the biological rationale it seems likely that this interaction is present in other cancer types as well. Technical challenges in this study were the high proportion of patients who had missing values for any of the data, and the possible non-linear relationships between the different variables and overall survival.

Part 2: estimating individual treatment effects

The study in **chapter 4** explored the possibility of estimating a treatment effect from medical images using a statistical technique called deep learning. The basis for this study is a combination of real CT images of lung tumors and simulated (artificial) outcome data. Due to the simulation, two characteristics of the tumors visible in the images were related to a simulated outcome: tumor size and tumor heterogeneity, measured as the total number of pixels in the tumor and variation in the intensity of pixels in the tumor respectively. The aim of the study was to estimate both patient prognosis and treatment effect based on the image of the tumor. Through the simulation, both the size and the heterogeneity of the tumor were correlated with the outcome. However, the size of the tumor was a collider. A collider distorts the estimation of the treatment effect if a model conditions on it. The only way to properly estimate the treatment effect in this situation was to ignore the collider (size of the tumor) in the prediction. At the same time, the whole image could not be ignored because then the information that the tumor heterogeneity holds about survival would be lost. The challenge with deep learning methods is that by virtue of the method, deep learning can discover all characteristics in an image that are related to the outcome. This is often attractive as it might discover new patterns that were previously unknown. As a downside, there is no direct way of knowing or controlling what patterns are used by the model. In this case we want the model to use the information on tumor heterogeneity but not tumor size, but when left unrestricted the deep learning model would use both characteristics, as both are correlated with the simulated outcome. To solve this problem, we devised a method in which a deep learning model was developed in two stages. In the first stage, both the outcome and the size of the tumor were predicted. In the second stage, the predicted tumor size information was shielded from the outcome prediction and the outcome was re-predicted based on all information from the image except tumor size. In this way it was possible to extract the important prognostic information from the image (tumor heterogeneity) and at the same time correctly estimate the treatment effect, by ignoring the collider tumor size. An important challenge in this study was connecting abstracted concepts such as tumor size and heterogeneity with the pixels in an image. The presented method worked very well for the simulated data, but needs to be extended to be more widely applicable.

In **chapter 5** we estimate the individual treatment effect for stage III non-small cell lung cancer patients. The treatment choice examined was the decision between chemotherapy and radiotherapy simultaneously (concurrent chemoradiation) or chemotherapy followed by radiotherapy (sequential chemoradiation). Previous randomized studies have shown that concurrent chemoradiation leads to better survival on average, but that not all patients are fit enough to undergo concurrent chemoradiation. The trade-off between sequential and concurrent chemoradiation must be made on an individual level. The aim of this study was to predict survival under both concurrent and sequential chemoradiation in order to support individual treatment decisions. The fundamental challenge was to estimate the causal effect of treatment on survival for the individual patient. In the data for this study, the choice between concurrent or sequential therapy was made based on standard clinical judgment, not randomization. Because patients in good overall fitness are more likely to receive concurrent treatment, the survival difference between patients with concurrent and sequential therapy in these patients does not equal the causal effect of the treatments. The treatment groups also differ in overall fitness. Better overall fitness leads to better survival, regardless of which treatment is chosen. Because overall fitness determines both treatment choice and outcome, overall fitness is called a 'confounder'. The causal effect of the treatments can be estimated in this population by grouping patients based on overall fitness and then comparing the difference in survival between concurrent and sequential chemoradiation at equal overall fitness. The crux of this study was that no direct measurement of overall fitness is available. The treating physician makes an implicit estimate of the patient's overall fitness based on many different signs and measurements. This includes objective measures such as age and performance score. However, there are also characteristics that are difficult to measure, such as how the patient looks, how the patient walks into the consultation room, how the patient's voice sounds, and so on. While these pieces of information are part of the physician's estimate of overall fitness, they are not recorded. Overall fitness is therefore not observed from the researcher's point of view and it is thus impossible to group patients on the basis of overall fitness. This means that overall fitness is an 'unobserved' confounder. Only variables such as age and performance score are known, but these are 'proxy' measures of overall fitness and not the real overall fitness as estimated by the treating physician. This unobserved confounder makes it impossible to make a fair comparison between patients treated with concurrent or sequential chemoradiation. This problem is not only present in stage III lung cancer but applies broadly health care.

There are existing methods for estimating treatment effects when only proxy measurements of unobserved confounders are available, but none of these methods were suited to our case. For this reason, we have developed a new method that makes it possible to estimate the causal treatment effect based on the proxy measurements. The method is called 'proxy-based individual treatment effect modeling in cancer' (PROTECT). In the PROTECT method, additional background knowledge is used to estimate the unobserved confounder "overall fitness" based on the available proxy measurements. For example, background knowledge says that patients with better overall fitness will also have a better performance score. Explicitly including background knowledge like this in the statistical model makes it possible to estimate the causal treatment effect. The application of PROTECT to the stage III non-small cell lung cancer patients yielded relevant findings. While conventional statistical methods seemed to overestimate the treatment effect, PROTECT's treatment effect estimate was in line with what was expected based on background knowledge. Due to the relatively limited number of patients included in the study (507), the statistical uncertainty about this estimate remained relatively high. It is important to further investigate the PROTECT method in other cancer types and in larger cohorts.

In **chapter 6** we evaluate offset models as a method for estimating individual treatment effects. Offset methods rely on the assumption that treatments have a constant and known 'relative' treatment effect, but that patients differ with respect to their 'baseline risk': the risk of an outcome if they would not be treated. For example, suppose that a certain cholesterol lowering drug reduces the 10-year risk of cardiovascular death by a factor of 2 for all patients. A 60-year-old male smoker with hypertension and raised cholesterol has a baseline risk of cardiovascular death of 40% and should expect a reduction in risk of 20% points. A 50-year-old female without hypertension has a baseline risk of under 1% and will have a less than 0.5% points reduction in risk. Given these different effects on an absolute probability scale, one may recommend the new cholesterol lowering drug to the 60-year-old male but not the 50-year-old female. The challenge with this approach is how to estimate the baseline risk. Oftentimes, randomized controlled trials are too small to estimate this risk on a granular level depending on many patient characteristics. Conversely, in historical observational patient data, some patients may have had the treatment whereas others have not. Because of confounding, the patients who did not get the treatment of interest are not a good reference population for estimating the baseline risk in the entire pre-treatment-decision population, see for example **Figure 3** in **chapter 7**. Offset models estimate the baseline risk for patients by accounting for the effects of historical treatments using a fixed 'offset term' that is based on an estimate of the relative treatment effect from prior randomized controlled trials. Variants of the offset method have been used for example to estimate the benefit of chemotherapy after surgical resection of a breast tumor. Some of these models are recommended for treatment decision support by current clinical guidelines. However, it is unknown if the assumption of a constant relative treatment effect underlying the offset method is indeed sufficient to account for unobserved confounding. In **chapter 6** we demonstrate that offset methods for binary outcomes do not estimate the ground truth baseline risk or treatment effect in the presence of confounding, but find that the resulting systematic error is low in most cases. Also, based on statistical considerations, we introduce a more refined way of using the estimate of the relative treatment effect by introducing a new constraint. We find that constrained offset models perform better than standard offset models, and are a defendable approach to individual treatment effect estimation whenever the assumption of a constant relative treatment effect is tenable.

The thesis concludes in **chapter 7** with a viewpoint article on the (dis)utility of outcome predictions for decision support in cancer. We note that predicting outcomes in itself is rarely the goal. Ultimately, the aim is to make a better treatment decision based on the prediction. The question "what is the expected survival, given that we know X about the patient?" is not the most important, but "what is the expected survival, *if we were to give treatment A or B*, given that we know X about the patient?". As shown in **chapter 4, chapter 5 and chapter 6**, answering the latter question is much more complex than the former question that is purely predictive. The crux of the matter lies in the causal nature of the treatment decision question: which treatment has a better survival as a causal effect? In **chapter 7** we emphasize the importance of this causal question and explain how much of the published research on predicting outcomes cannot contribute to answering this question because the causal nature of the question is not acknowledged. In the article we then discuss what is needed to answer this important question properly. We conclude with a call for a better understanding of the methodological knowledge needed to answer causal questions and for more research that can provide an answer to the question that really matters in clinical practice. To achieve the goal of a care that is increasingly tailored to the individual patient, it is crucial to strive for research of the highest methodological standard.

CHAPTER 10

Nederlandse Samenvatting

Met een geschatte 1.761.007 overlijdens per jaar is longkanker de belangrijkste oorzaak van kanker gerelateerde sterfte wereldwijd. In Nederland is longkanker eveneens de grootste oorzaak van kanker gerelateerde sterfte en overlijden jaarlijks ongeveer 14.000 mensen aan longkanker. Een bepalende factor voor de kans op overleving is het stadium van de ziekte bij diagnose. Als de longkanker in een vroeg stadium wordt gevonden kan het vaker genezen worden. Als de ziekte al verspreid is naar andere organen is genezing in de regel niet meer mogelijk. Op basis van internationale afspraken wordt longkanker in vier stadia verdeeld naar gelang hoe ver de ziekte verspreid is. Voor patiënten met het vroegste ziektestadium, stadium I, is de vijfjaarsoverleving is 53%. Dat betekent dat van alle patiënten die bij diagnose van de longkanker in ziektestadium I zijn, na vijf jaar nog 53% in leven is. Voor stadium II is de vijfjaarsoverleving 38%, voor stadium III is dit 17% en voor stadium IV slechts 3%. Hoewel het ziektestadium dus veelzeggend is voor overleving, is er ook binnen een stadium nog veel spreiding in overleving. Zo overlijdt 25% van de stadium II patiënten binnen een jaar na diagnose, maar een andere 25% van de patiënten leeft na 10 jaar nog. Op basis van hoe de cellen van een longtumor er onder de microscoop uit zien wordt longkanker in twee grote groepen verdeeld: kleincellige longkanker en niet-kleincellige longkanker. Ongeveer 85% van de longkankers in Nederland zijn niet-kleincellig en dit proefschrift is toegespitst op deze variant.

Voorspellen van overleving

Gezien deze grote spreiding in overleving is het toespitsen van de verwachte levensduur op de individuele patiënt een lang bestaand onderzoeksdoel. **Deel 1** van dit proefschrift is gewijd aan het voorspellen van overleving voor longkankerpatiënten op basis van de gegevens die ten tijde van de diagnose bekend zijn. Een deel van de spreiding in overleving is gebaseerd op het toeval. Er zijn bekende risicofactoren die de kans op het krijgen van longkanker vergroten, zoals roken. Toch is het krijgen van longkanker uiteindelijk het resultaat van een interne en willekeurige mutatie. Nadat een longtumor ontstaan is, blijven er willekeurige mutaties optreden die het verdere ziektebeloop en de gevoeligheid van de tumor voor eventuele behandelingen bepalen. Naast de longkanker is de overleving van een patiënt ook aan externe willekeurigheid onderhevig, zoals het krijgen van infectieuze ziekten, slachtoffer worden van een ongeval of oneindig veel meer mogelijke factoren. Dit maakt het fundamenteel onmogelijk om op het moment van diagnose met hoge nauwkeurigheid de overleving van een patiënt te voorspellen. Naast deze willekeurigheden zijn er echter ook eigenschappen van de longtumor en van de patiënt zelf die van belang zijn voor overleving en die wel gemeten kunnen worden op het moment van de diagnose. Deze eigenschappen kunnen worden opgedeeld in twee groepen: eigenschappen van de tumor en eigenschappen van de patiënt. In **hoofdstuk 2** van dit proefschrift worden eigenschappen van de tumor die voorspellend zijn voor overleving nader onderzocht. In **hoofdstuk 3** wordt een studie gepresenteerd die naar eigenschappen van de patiënt kijkt, specifiek eigenschappen die gerelateerd zijn aan de hoeveelheid spiermassa van de patiënt.

Het belang van beeldvorming

Als de diagnose longkanker gesteld is, maar voordat er een keuze gemaakt wordt voor een eventuele behandeling, is het van belang om de longtumor zo goed mogelijk in beeld te brengen. Naast het verkrijgen van weefsel van de tumor voor microscopisch onderzoek is radiologische beeldvorming van groot belang. Beeldvorming bij longkanker is met name gebaseerd op "computed tomography scans", CT-scans met intraveneus jodiumhoudend contrast. Bij een CT-scan worden veel opeenvolgende röntgenfoto's genomen die telkens een paar millimeter opgeschoven zijn. Elke afbeelding van een CT-scan is een doorsnede van een patiënt op een bepaalde hoogte, gemeten vanaf de tenen naar de kruin. Door veel doorsnedes te maken op korte afstand van elkaar kan er door de patiënt 'gescrolled' worden en ontstaat er beeld van de 3D anatomie. De röntgenfoto's geven verschillen in radiodichtheid weer als verschillende lichtintensiteiten in het beeld. Radiodichtheid is nagenoeg recht evenredig met dichtheid van massa (gewicht gedeeld door volume). Omdat longen luchthoudend zijn en dus een veel lagere dichtheid hebben dan een longtumor resulteert dit in een helder zichtbaar contrast op CT beelden. Op CT-scans is te zien hoe groot een tumor is, hoe de tumor gesitueerd is ten opzichte van andere organen, of de tumor mogelijk ingroeit in andere organen en of er op meerdere plekken in het lichaam mogelijke uitzaaiingen zijn van de tumor. Daarnaast is te zien of de tumor een gelijkmatige dichtheid heeft, of dat er misschien regio's in de tumor zijn met een verschillende dichtheid. Als deze informatie over een tumor gekwantificeerd wordt in metingen kan worden onderzocht of deze metingen samenhangen met overleving. Deze kwantitatieve metingen zijn mogelijke 'prognostische factoren'. Een voor de hand liggende vraag is bijvoorbeeld of de grootte van een tumor samenhangt met een slechtere overleving. Gezien de wijdverbreidheid van CT-scans in longkankerzorg is er veel onderzoek gedaan naar hoe bepaalde eigenschappen van de longtumor, zoals zichtbaar op CT-scans, samenhangen met overleving. Een belangrijk gegeven bij het zoeken naar nieuwe prognostische factoren is wat de nieuwe informatie toevoegt aan informatie die al standaard voorhanden is tijdens het klinisch proces. Bekende eigenschappen van een patiënt die belangrijk zijn voor overleving zijn leeftijd, algemene gezondheid en de aanwezigheid van gewichtsverlies. Een veelgebruikte meting van algemene gezondheid is "performance score", een semi-kwantitatieve meting van hoe goed een patiënt in staat is om voor zichzelf te zorgen.

Overzicht van de literatuur

In **hoofdstuk 2** presenteren we een studie die de literatuur samenvat over prognostische factoren meetbaar op CT-scans voor stadium III niet-kleincellig longkanker. De studie was uitgevoerd met een uitgebreide systematische zoekstrategie. Na selectie op relevantie voor de onderzoeksvraag werden er 65 artikelen geïncludeerd die samen 26 unieke prognostische factoren beschrijven en gezamenlijk 144.513 patiënten geïncludeerd hadden. Alle artikelen werden beoordeeld op kwaliteit om in te schatten hoe betrouwbaar de gepresenteerde resultaten zijn. Er was een grote variatie in kwaliteit, en slechts 4 studies vergeleken de toegevoegde waarde van potentiële nieuwe prognostische factoren gemeten op CT-scans met de standaard beschikbare prognostische factoren. Deze samenvatting van de literatuur wijst erop dat het totale volume van de primaire tumor, lymfeklier uitzaaiingen, de tumor diameter, het volume van lymfeklier uitzaaiingen en de aanwezigheid van longvliesvocht gerelateerd zijn aan overleving. Door de manier waarop de analyses zijn gedaan is het niet duidelijk hoe meerdere prognostische factoren samen gerelateerd zijn met overleving. Het kan zijn dat sommige prognostische factoren elkaar versterken of juist verzwakken, maar dat wordt niet beantwoord door de gebruikte analysemethoden. Aanbevelingen die in deze studie gedaan worden zijn dat toekomstige studies over prognostische factoren beter hun methode en resultaten moeten rapporteren, en potentiële nieuwe prognostische factoren moeten vergelijken met bekende prognostische factoren.

Spier volume en dichtheid zijn belangrijk voor overleving

Naast eigenschappen van de tumor zijn eigenschappen van de patiënt ook gerelateerd aan overleving. Eerder genoemd is dat performance score een maat is voor algemene gezondheid. Een andere meting die gerelateerd is aan gezondheid is de body-mass index (BMI). Om de algemene gezondheid nog beter in kaart te brengen zijn er verscheidene mogelijke metingen uit te voeren op CT-scans. Zo kan er op CT-scans bijvoorbeeld gekeken worden naar de verspreiding van vet en naar de hoeveelheid spierweefsel van een patiënt. Een veelgebruikte maat om op CT-scans de hoeveelheid spiervolume van een patiënt te meten is door op het niveau van de 3^e rugwervel te kijken hoe groot het spieroppervlak van de psoas-spier is. Dit oppervlak correleert sterk met de totale hoeveelheid spiervolume in het gehele lichaam. Door dit psoas spieroppervlak te delen door het kwadraat van de lengte van een patiënt krijgt men de ‘psoas muscle index’ (PMI), een soort spier-analoog van de BMI. In veel kankersoorten is er een verband aangetoond tussen PMI en overleving. Patiënten met een hogere PMI (meer spier) leven gemiddeld langer, vermoedelijk omdat ze in beter algemene conditie zijn en/of omdat de tumor een minder hoge tol heeft geëist op het lichaam op het moment van de diagnose. Naast de hoeveelheid spier is de kwaliteit van de spier ook van belang. Door verscheidene processen, waaronder veroudering, immobiliteit en ontstekingsreacties gerelateerd aan kanker, kan spierweefsel langzaam vervangen worden door vetweefsel. Hierdoor wordt de kwaliteit van het spierweefsel aangetast. Vetweefsel heeft een lager dichtheid dan spierweefsel. De dichtheid van een spier zoals gemeten op een CT-scan is dus een maat voor de hoeveelheid vetinfiltratie in die spier en een maat van spierkwaliteit.

De gemiddelde dichtheid van de psoas-spier gezien op een CT-scan op het niveau van de 3^e rugwervel is een gestandaardiseerde meting die ‘psoas muscle radiodensity’ heet (PMD). In **hoofdstuk 3** presenteren we een studie die de samenhang PMI en PMD met overleving in niet-kleincellig longkanker bestudeert. Er zijn meerdere studies die de samenhang tussen spier kwantiteit, spier dichtheid en overleving hebben onderzocht in longkankerpatiënten, maar de studies hebben conflicterende resultaten. Als mogelijke verklaring voor de tegenstrijdigheden presenteerden wij de hypothese dat de samenhang tussen PMI en overleving sterker is als de spierdichtheid, gemeten in PMD, voldoende is. De biologische motivatie achter deze hypothese is dat de kwantiteit van spier (PMI) alleen samenhangt met betere overleving als de kwaliteit, gemeten met de spierdichtheid PMD, voldoende is. In statistische termen betekent dit dat er een statistische interactie is tussen PMI en PMD en overleving. Als de hypothese waar is, en als de patiënten in de eerder gepubliceerde studies gemiddeld van elkaar verschillen wat betreft PMD, zou dit kunnen verklaren dat de gevonden samenhang tussen PMI en overleving verschilt per studie. Om deze vraag te beantwoorden hebben we een grote groep van 2480 niet-kleincellig longkankerpatiënten verzameld die behandeld zijn bij de radiotherapieafdeling van het Universitair Medisch Centrum Utrecht. Met een geautomatiseerd computeralgoritme gebaseerd op een techniek die ‘deep learning’ heet, is bij deze patiënten een meting van de PMI en PMD gedaan. De samenhang tussen PMI, PMD en overleving werd geanalyseerd in de context van bekende eigenschappen van de tumor (histologisch subtype) en de patiënt (leeftijd, geslacht, performance score en BMI). Er was duidelijk statistisch bewijs voor onze hypothese. Dit betekent dat toekomstige studies over de associatie tussen spierhoeveelheid en algehele overleving rekening moeten houden met het effect dat spierdichtheid heeft op de associatie tussen spierhoeveelheid en algehele overleving. Hoewel onze studie alleen werd uitgevoerd bij patiënten met niet-kleincellig longkanker, lijkt het, gezien de biologische achtergrond van de hypothese, waarschijnlijk dat deze interactie ook bij andere kankertypes aanwezig is. Technische uitdagingen bij deze studie waren het grote aandeel van patiënten die missende waarden hadden voor een van de gegevens, en de mogelijke niet-lineaire verbanden tussen de verschillende variabelen en het risico op overlijden.

Individueel behandeleffect schatten

In **deel 2** van dit proefschrift hebben we de onderzoeksvraag verschoven van het voorspellen van overleving naar het beantwoorden van een nog belangrijkere vraag. De meest relevante vraag in de klinische praktijk is: “Wat zou de verwachte overleving zijn, als we behandeling A of B zouden geven, gegeven dat we eigenschap X over deze patiënt weten”. Deze vraag gaat over het *individuele behandeleffect* en gaat er dus van uit dat niet elke behandeling voor elke tumor en patiënt even goed zal werken. Deze vraag is een niveau complexer dan alleen overleving voorspellen. Dit komt omdat het gaat over het *oorzakelijke* effect van de behandeling. Oorzakelijke effecten (causaliteit) zijn niet hetzelfde als *(statistische) correlaties*. Er zijn veel statistische correlaties die niet op een oorzakelijk effect berusten. Zo is er bijvoorbeeld een sterke correlatie tussen de hoeveelheid mozzarella consumptie per persoon in de Verenigde Staten en het aantal civiel ingenieur doctoraten in de periode tussen 2000 en 2009 (voor statistisch geletterden, de r-coëfficiënt is 0.96, een zeer sterke correlatie). Zie de website https://www.tylervigen.com/spurious-correlations voor meer voorbeelden. Om oorzakelijke verbanden aan te tonen zijn ofwel experimenten nodig ofwel gedetailleerde kennis van hoe de gegevens precies verzameld zijn en specifiek *waarom* bepaalde behandelingen historisch gegeven zijn.

Verstorende signalen filteren

De studie in **hoofdstuk 4** onderzocht de mogelijkheid om een behandeleffect te schatten op basis van medische beeldvorming terwijl er gebruik gemaakt wordt van een computeralgoritme dat ‘deep learning’ heet. Zoals we in **hoofdstuk 2** en **hoofdstuk 3** gezien hebben bevat een CT-scan informatie over zowel de tumor als de patiënt. Deze informatie hangt mogelijk samen met de effectiviteit van verschillende behandelingen voor een individuele patiënt. Door deze informatie te gebruiken voor het schatten van een individueel behandeleffect kan de behandeling voor patiënten gepersonaliseerd worden. De gebruikte techniek ‘deep learning’ is hierbij aantrekkelijk omdat dit algoritme zelf leert om de relevante patronen in afbeeldingen te herkennen. Conventioneel onderzoek is gebaseerd op gerichte metingen van karakteristieken van de beeldvorming die de onderzoeker zelf heeft bedacht, zoals de grootte van de tumor, de hoeveelheid spieroppervlak enzovoorts. Aangezien de onderzoeker nooit alles kan weten van wat er belangrijk is om te meten, kan het zijn dat er belangrijke informatie gemist wordt. Het voordeel van deep learning is dat dit algoritme *zelf* leert welke patronen in beelden relevant zijn voor wat er voorspeld moet worden. Dit doet het algoritme door te leren van voorbeelden. In een ‘trainingset’ worden combinaties aangeboden van beelden (bijvoorbeeld uit een CT-scan) en uitkomsten (bijvoorbeeld de overleving van een patiënt). Met een optimalisatietechniek leert het algoritme om de meest voorspellende patronen in het beeld te herkennen. Een nadeel aan deep learning is dat het model een relatieve ‘black box’ is. Het is niet mogelijk om direct te achterhalen op basis van welke patronen het model een bepaalde voorspelling doet.

In de simulatiestudie van **hoofdstuk 4** bestudeerden we een belangrijk probleem dat gerelateerd is aan de ‘black-box’ voorspellingen van deep learning. De basis voor deze studie is een combinatie van echte CT afbeeldingen van longtumoren en gesimuleerde (kunstmatige) uitkomstdata. Door de simulatie waren er twee eigenschappen van de tumoren zichtbaar op de afbeeldingen die gerelateerd waren aan een gesimuleerde uitkomst: overleving. Het ging hierbij om de omvang van de tumor en de heterogeniteit van de tumor. Heterogeniteit was gedefinieerd als de variatie van de intensiteit van de pixels van de tumor. Tumoren met subregio’s van verschillende dichtheid hebben een hogere heterogeniteit dan tumoren die uit één homogeen weefseltype bestaan. Het doel van de studie was om zowel de prognose van de patiënt als het behandeleffect te schatten, op basis van de afbeelding van de tumor. Door de simulatie waren zowel de omvang als de heterogeniteit van de gecorreleerd met de uitkomst. Echter, de omvang van de tumor was een ‘collider’. De precieze definitie van een collider is hier niet belangrijk, maar het netto-effect van een collider is dat een collider de schatting van het behandeleffect verstoort. De enige manier om in deze situatie het behandeleffect goed te schatten is door de collider (omvang van de tumor) te negeren in de voorspelling. Tegelijkertijd kon niet de hele afbeelding genegeerd worden omdat dan de informatie die de tumor heterogeniteit over overleving had verloren zou gaan. Om dit probleem op te lossen hebben we een methode bedacht waarbij een deep learning model in twee stadia werd ontwikkeld. In het eerste stadium werden zowel de uitkomst als de omvang van de tumor voorspeld. In de tweede fase werd de voorspelde informatie over tumor omvang afgeschermd van de voorspelling van de uitkomst en werd de uitkomst opnieuw voorspeld op basis van alle informatie uit de afbeelding, behalve de omvang van de tumor. Op deze manier was het mogelijk om de belangrijke prognostische informatie uit het plaatje te extraheren (tumor heterogeniteit) en tegelijkertijd het behandel effect correct te schatten, door de collider ‘tumor omvang’ te negeren. Een belangrijke uitdaging bij deze studie was het verbinden van abstracte begrippen zoals tumor omvang en heterogeniteit en de pixels in een afbeelding. De gepresenteerde methode werkte erg goed voor de gesimuleerde data, maar moet nog uitgebreid worden om breder toepasbaar te zijn.

Corrigeren voor latente confounders

In **hoofdstuk 5** presenteren we wederom een studie waarin het individuele behandeleffect centraal staat, ditmaal toegepast op stadium III niet-kleincellig longkankerpatiënten. De behandelkeuze die onderzocht werd was de afweging tussen chemotherapie en radiotherapie tegelijkertijd (concurrente chemoradiatie) of chemotherapie gevolgd door radiotherapie (sequentiële chemoradiatie). Uit eerdere gerandomiseerde studies is gebleken dat concurrente chemoradiatie tot betere overleving leidt, gemiddeld genomen, maar dat niet alle patiënten fit genoeg zijn om concurrente chemoradiatie te ondergaan. De afweging tussen sequentiële en concurrente chemoradiatie moet op een individueel niveau gemaakt worden. Het doel van deze studie was om de overleving te voorspellen onder zowel concurrente als sequentiële chemoradiatie om op deze manier de individuele afweging te ondersteunen. Hierbij was het van cruciaal belang om het oorzakelijke effect van de behandeling op overleving voor de individuele patiënt te schatten. De gegevens voor deze studie kwamen van patiënten die in het Universitair Medisch Centrum Utrecht behandeld waren met radiotherapie. De keuze tussen concurrent of sequentiële therapie werd gemaakt op basis van standaard klinische afwegingen, niet door randomisatie. Omdat patiënten die in goede algemene gezondheid zijn vaker concurrente behandeling krijgen, is het overlevingsverschil tussen patiënten met concurrente en sequentiële therapie in deze patiënten niet gelijk aan het oorzakelijke effect van de behandeling. De patiënten verschillen namelijk ook met betrekking tot algemene gezondheid. Betere algemene gezondheid leidt tot betere overleving, onafhankelijk van welke behandeling er wordt gekozen. Omdat algemene gezondheid zowel bepalend is voor de behandelkeuze als de uitkomst wordt algemene gezondheid een 'confounder' genoemd. De enige manier om het juiste behandeleffect in deze populatie te schatten is door patiënten te groeperen op basis van gezondheid en dan het verschil in overleving tussen concurrente en sequentiële chemoradiatie *bij gelijke algemene gezondheid* te vergelijken. De crux bij deze studie was dat er geen goede meting van algemene gezondheid beschikbaar is. De behandelend arts maakt een impliciete schatting van de gezondheid van de patiënt op basis van veel verschillende gegevens. Het gaat hierbij om objectieve metingen zoals leeftijd en 'performance score'. Er spelen echter ook factoren mee die moeilijk meetbaar zijn, zoals hoe de patiënt er uit ziet, hoe de patiënt de spreekkamer in loopt, hoe de stem van de patiënt klinkt, enzovoorts. Hoewel deze gegevens deel uitmaken van de inschatting van de arts zijn ze niet beschikbaar voor onderzoek. Dit betekent dat de algemene gezondheid vanuit het oogpunt van de wetenschapper niet geobserveerd is en het dus onmogelijk is om patiënten correct te groeperen op basis van gezondheid. We spreken hier van een *latente confounder*. Alleen variabelen zoals leeftijd en performance score zijn bekend, maar dit zijn 'proxy' metingen van gezondheid en niet de echte algemene gezondheid zoals de behandelend arts het inschat. Door deze latente confounder is het onmogelijk om een eerlijke vergelijking te maken tussen patiënten die met concurrente of sequentiële chemoradiatie zijn behandeld. Dit probleem speelt zich niet alleen bij stadium III longkanker af maar is wijdverbreid in de gezondheidszorg. Tegelijkertijd is er een grote vraag naar geïndividualiseerde behandeladviezen op basis van historische niet-experimentele gegevens. Hierom hebben we een methode ontwikkeld die het mogelijk maakt om op basis van de proxy metingen toch nog het oorzakelijke behandeleffect te schatten. De methode heet 'proxy based individual treatment effect modeling in cancer' (PROTECT). Bij de PROTECT methode wordt aanvullende achtergrondkennis gebruikt om de latente confounder "algemene gezondheid" te schatten op basis van de aanwezige proxy metingen. De achtergrondkennis is bijvoorbeeld dat patiënten met betere gezondheid ook een betere performance score zullen hebben. Door deze achtergrondkennis expliciet op te nemen in het statistische model wordt het mogelijk om het oorzakelijke behandeleffect te schatten. De toepassing van PROTECT op de stadium III niet-kleincellig longkankerpatiënten leverde relevante bevindingen op. Waar conventionele statistische methoden een overschatting leken te geven van het behandeleffect, was de behandeleffectschatting van PROTECT in lijn met wat er verwacht werd op basis van achtergrondkennis. Door het relatief beperkt aantal patiënten dat geïncludeerd was in de studie (507) bleef de statistische onzekerheid over deze schatting relatief groot. Het is van belang om de PROTECT methode bij andere kankersoorten en in grotere populaties verder te onderzoeken.

Van relatief naar absoluut behandeleffect

In **hoofdstuk 6** evalueren we zogenaamde 'offsetmethoden' voor het schatten van individuele behandeleffecten. Offsetmethoden gaan op een andere manier om met latente confounders, door de aanname te maken dat behandelingen een constant en bekend 'relatief' behandeleffect hebben, maar dat patiënten verschillen met betrekking tot hun 'basisrisico': het risico op een uitkomst als ze niet behandeld zouden worden. Als voorbeeld, stel dat een bepaald cholesterolverlagend medicijn het 10-jaarsrisico op cardiovasculaire sterfte met een factor 2 vermindert. Een 60-jarige mannelijke roker met hypertensie en verhoogd cholesterol heeft een basisrisico op cardiovasculair overlijden van 40% en zou een risicoreductie van 20% punten hebben. Een 50-jarige vrouw zonder hypertensie heeft een basisrisico van minder dan 1% en zal een risicoreductie van 0,5% punten hebben. Gezien deze sterk verschillende effecten in absolute procentpunten, kan men het nieuwe cholesterolverlagende medicijn bijvoorbeeld aanbevelen aan de 60-jarige man, maar niet aan de 50-jarige vrouw. De uitdaging bij deze benadering is hoe het basisrisico kan worden ingeschat. Vaak zijn gerandomiseerde onderzoeken te klein om dit risico op een granulair niveau in te schatten, afhankelijk van de kenmerken van de patiënt. Omgekeerd, in historische observationele patiëntgegevens kunnen sommige patiënten de behandeling hebben gehad, terwijl anderen dat niet hebben gedaan. Vanwege confounders zijn de patiënten die de behandeling niet hebben gekregen geen goede referentiepopulatie voor het schatten van het basisrisico in de gehele populatie voorafgaand aan de behandel beslissing. Offsetmodellen schatten het basisrisico voor patiënten door rekening te houden met de effecten van historische behandelingen met behulp van een vaste 'offsetterm' die is gebaseerd op een schatting van het relatieve behandelingseffect uit eerdere gerandomiseerde onderzoeken. Varianten van de offsetmethode zijn bijvoorbeeld gebruikt om het voordeel van chemotherapie na chirurgische resectie van een borsttumor in te schatten. Deze modellen worden aanbevolen voor ondersteuning van de behandelbeslissing door de huidige klinische richtlijnen. Het is echter niet bekend of de aanname van een constant relatief behandelingseffect die ten grondslag ligt aan de offsetmethode inderdaad voldoende is om het effect van latente confounders teniet te doen. In **hoofdstuk 6** laten we zien dat offset-methoden voor binaire uitkomsten niet het basisrisico of het behandelingseffect schatten in de aanwezigheid van confounders, maar dat de resulterende systematische fout in de meeste gevallen laag is. Ook introduceren we, op basis van statistische overwegingen, een meer verfijnde manier om de schatting van het relatieve behandelingseffect te gebruiken door een nieuwe restrictie te introduceren. We vinden dat offsetmodellen met de nieuwe restrictie beter presteren dan standaard offsetmodellen, en een verdedigbare benadering zijn voor het schatten van individuele behandelingseffecten wanneer de aanname van een constant relatief behandelingseffect houdbaar is.

De beste behandeling kiezen, een causale taak

Het proefschrift sluit af in **hoofdstuk 7** met een ‘viewpoint’ artikel over de waarde het voorspellen van uitkomsten voor behandelbeslissingen in kanker. De centrale stelling is dat het voorspellen van uitkomsten op zichzelf slechts zelden het ultieme doel is. Uiteindelijk is het de bedoeling om op basis van de voorspelling een betere behandelkeuze te maken. De vraag “wat is de verwachte overleving, gegeven dat we X over de patiënt weten?” is niet de belangrijkste, maar “wat is de verwachte overleving, als we behandeling A of B zouden geven, gegeven dat we X over de patiënt weten?”. Zoals in **hoofdstuk 4, 5** en **6** is aangetoond is het beantwoorden van deze vraag methodologisch veel complexer dan de vraag die puur voorspellend is. De crux zit in het oorzakelijke karakter van de behandelbeslissingvraag: welke behandeling heeft als oorzakelijk effect een betere overleving? In **hoofdstuk 7** benadrukken we het belang van deze vraag en leggen we uit hoe een groot deel van het gepubliceerde onderzoek over het voorspellen van uitkomsten bij kanker niet kan bijdragen aan het beantwoorden van deze vraag omdat het oorzakelijke karakter van de vraag niet goed wordt meegenomen in de opzet van de studie en de uiteindelijke analyse. In het artikel bespreken we vervolgens wat er nodig is om deze belangrijke vraag *wel* goed te beantwoorden. We sluiten af met een oproep voor een betere verdieping in de methodologische kennis rondom oorzakelijke vragen en voor meer onderzoek dat een antwoord kan geven op de vraag waar het in de kliniek echt om gaat. Om het doel te behalen van een zorg steeds beter toegespitst wordt op de individuele patiënt is het cruciaal om onderzoek van de hoogste methodologische standaard na te streven.

List of publications

Publications included in this thesis

van Amsterdam WAC, Verhoeff JJC, de Jong PA, Leiner T, Eijkemans MJC. Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning. npj Digital Medicine. 2019 Dec 10;2(1):1–6.

van Laar M, van Amsterdam WAC, van Lindert ASR, de Jong PA, Verhoeff JJC. Prognostic factors for overall survival of stage III non-small cell lung cancer patients on computed tomography: A systematic review and meta-analysis. Radiotherapy and Oncology. 2020 Oct 1;151:152–75.

van Amsterdam WAC, Verhoeff JJC, Harlianto NI, Bartholomeus GA, Puli AM, de Jong PA, et al. Individual treatment effect estimation in the presence of unobserved confounding using proxies: a cohort study in stage III non-small cell lung cancer. Sci Rep. 2022 Apr 7;12(1):5848.

van Amsterdam WAC, Harlianto NI, Verhoeff JJC, Moeskops P, de Jong PA, Leiner T. The association between muscle quantity and overall survival depends on muscle radiodensity: a cohort study in non-small cell lung cancer patients. Journal of Personalized Medicine (2022) 12(7), 1191.

van Amsterdam WAC, Ranganath R. Conditional Average Treatment Effect estimation with Treatment Offset Models. submitted. 2022;

van Amsterdam WAC, de Jong PA, Suijkerbuijk KPM, Verhoeff JJC, Leiner T, Ranganath R. Decision making in cancer: causal questions require causal answers. submitted. 2022;

Other publications

van Amsterdam WAC, Blankestijn PJ, Goldschmeding R, Bleys RLAW. The morphological substrate for Renal Denervation: Nerve distribution patterns and parasympathetic nerves. A post-mortem histological study. Annals of Anatomy - Anatomischer Anzeiger. 2016 Mar;204:71–9.

Bastiaannet R, van Roekel C, Smits MLJ, Elias SG, van Amsterdam WAC, Doan D, et al. First Evidence for a Dose–Response Relationship in Patients Treated with 166Ho Radioembolization: A Prospective Study. Journal of Nuclear Medicine. 2020 Apr 1;61(4):608–12.

Dankwoord

Dit proefschrift was nooit tot stand gekomen zonder de hulp en bijstand van velen, waarvan ik er hier enkelen wil noemen.

Ten eerste mijn promotoren en copromotoren.

Prof. Dr. Pim de Jong, ik ben je zeer erkentelijk voor de begeleiding die je me de afgelopen jaren hebt gegeven. Je bent een harde werker, zowel in de kliniek, de wetenschap, als het management. Ik ken geen andere coauteur die zo snel een inhoudelijke reactie geeft op stukken. Je loopt over van ideeën voor wetenschappelijke projecten. Ik ben blij dat je me begeleid hebt.

Prof. Dr. Tim Leiner, toen ik op zoek was naar een machine learning project binnen het UMC Utrecht (voordat dit 'hip' was), was jij de eerste met wie ik een 'click' had over dit onderwerp. Jij blinkt uit in (wetenschappelijke) ondernemendheid. Jouw enthousiasme voor techniek en innovatie, je brede netwerk, je ideeën en ambities zijn stuk voor stuk inspirerend voor me geweest.

Dr. Joost Verhoeff, door jouw domeinexpertise en praktijkervaring in longkanker hebben wij veel tijd van overleg gehad, dit ik altijd als prettig heb ervaren. Verder ben je een enthousiaste begeleider met een goed oog voor samenwerking en organisatie, waarin ik ook het nodige van je heb mogen leren. Daarnaast vond ik het erg leuk hoe we samen wetenschapsstudenten hebben begeleid.

Rajesh Ranganath, PhD. We got in touch early during my PhD because I was looking for researchers with expertise in machine learning and causal inference. Even though I had a limited understanding of my proposed project at the time, you invited me to your lab at the Courant Institute of Mathematical Sciences at New York University. Despite that it was just for 3 months, our time there was truly formative for me. The work atmosphere in that superstar city with superstar researchers was certainly the best working experience I had, despite risking my life twice a day by commuting between the office in Manhattan and our apartment in Brooklyn by bike. This visit confirmed my ambition to become a full-time researcher. I am glad our project took so long to complete and that we're still collaborating to this day.

Hartelijk dank aan de beoordelingscommissie van mijn proefschrift **Prof. Dr. Paul van Diest, Prof. Dr. Daniel Oberski, Prof. Dr. Rolf Groenwold, Prof. Dr. Miriam Koopman en Dr. Joop de Lange.**

Prof. Dr. René Eijkemans, jij raakte betrokken bij mijn promotie als 'tutor' van mijn epidemiologie-project, maar al snel bloeide dit uit tot een wetenschappelijke vriendschap. In de tijd dat ik me steeds verder verdiepte in statistiek, machine learning en causaliteit, waren onze wekelijkse gesprekken een baken in de storm voor mij. Jouw vragen en suggesties die voortkomen uit je brede kennis en ervaring zijn van groot belang geweest voor mijn ontwikkeling als wetenschapper en voor mijn promotie. Ik kijk met veel dankbaarheid terug op onze samenwerking.

Docenten van de epidemiologie master, waaronder Rebecca Stellato, Kas Kruithof, Rolf Groenwold, Sjoerd Elias, Rene Eijkemans, Carl Moons, Maarten van Smeeden, Thomas Debray, Herbert Hoijtink en anderen, jullie hebben een belangrijke bijdrage geleverd aan het fundament van mijn wetenschappelijke carrière.

Collega-studenten van de epidemiologiemaster, met name Aernoud Fiolet en Koos Korsten, wat is het mooi en inspirerend om zo'n vormende tijd samen door te gaan!

Medewerkers van de Radiotherapie, waaronder Alexis Kotte, Gijs Bol en Gery Thijsseling, zonder jullie hulp waren de meeste hoofdstukken van dit proefschrift er nooit geweest.

Wetenschapsstudenten die ik heb begeleid, met name **Myra van Laar, Netanja Harlianto, Gijs Bartholomeus en Sjors Witteveen**. Het was stimulerend en leerzaam voor mij om jullie te mogen begeleiden tijdens een deel jullie opleiding. Enkelen van jullie hebben terecht een plek in de auteurslijst van gepubliceerde artikelen in dit proefschrift.

Collega's van de Radiologie, met name Robbert, Bianca, Ahmed, Caren, Esmee, Floor, Frans, Jonas, Josanne, Justine, Liselore, Marcia, Margo, Mimount, Nienke, Rens, Sander, Sarah, Wieke. Met veel plezier kijk ik terug op onze lunchbesprekingen bij epi-rad, en de nodige collegiale borrels (hoewel er daar dankzij de coronaepidemie wel wat weinig van zijn geweest).

Mede onderzoekers uit de **(wal)vissenkom** Robbert, Jonas, Nienke, Floor en Niels. Jullie gaven het werken kleur. Onze koffies, wandelingen, en later ook dagopeningen gaven noodzakelijke opfrissing wanneer we het werk ons weer eens teveel in beslag nam.

PREMIUM collega's Karijn, Paul, Mitko, Josien, Rens en Belle. Dank voor de prettige samenwerking en dat ik deel uit kon maken van dit gave project.

Co-workers from CIMS-NYU: Aahlad, I greatly enjoyed our chats and collaboration. Mark, Mukund and Xintian, thank you for letting me be part of the team and helping me while I was at NYU (and after).

Suerman collegae, wat heb ik een geluk gehad dat mijn promotie aangevuld werd met suerman masterclasses en collega's. Dankzij onze activiteiten, intervisies en borrels heb ik me veel breder als professional en persoon kunnen ontwikkelen. In het bijzonder dank aan **Lisan van Os** voor de organisatie, en beschikbaarheid voor hulp waar nodig.

Collega's van het imaging science institute, Ivana Isgum, Bob de Vos, Jelmer Woltering, Majd Zreik. Dank voor jullie hulp toen ik de eerste stappen in Deep Learning voor medische beeldvorming aan het zetten was.

Oud collega's van de anatomie afdeling, met name Prof. Dr. Ronald Bleys en Suzanne, bij jullie heb ik mijn eerste stappen in de medische wetenschap gezet. Dank voor het vertrouwen en de begeleiding daarbij.

Dank aan alle personen en organisaties die financieel hebben bijgedragen aan deze promotie. De Suermancommissie van het UMC Utrecht, stichting de Drie Lichten, Girard de Mielet van Coehoorn stichting, NVIDIA, en het PREMIUM project (gefinancierd door ZonMW).

Het fundament. Vrienden van Archos (Merijn, Marijn, Glenn, Daniel, Rikjan), onze borrels en jullie relativeringen en aanmoedigingen hielpen mij in het rechte spoor te blijven. Vrienden van Lepelenburg, met name Jordi, Paul, Adolfo, meneer Pauw en Tato(†), jullie hielpen mij de echte waarde van werk te ontdekken en me te ontwikkelen als mens en professional.

Werner, onze vriendschap is en blijft voor mij onmisbaar. Jouw aanmoediging en soms relativering, jouw geduld om mijn monologen over mijn (wetenschappelijke) fascinaties aan te horen, ik ben een gezegend mens.

Schoonfamilie Maarten en Ineke, Corné en Kim, Susan en Soeneeth, al vele jaren vind ik bij jullie een tweede thuis.

Pap en mam, jullie hebben me altijd gestimuleerd om het best uit mezelf te halen. Jullie waren en blijven een fundament voor mij om op te bouwen. Ik besef me vaak, maar niet vaak genoeg, hoeveel geluk ik heb gehad bij jullie te mogen opgroeien, en nog steeds geniet ik van jullie steun. Marleen en Koen, Robin en Stijn, bedankt voor jullie steun en interesse.

Tenslotte mijn gezin, kinderen Vesper en Doris, het zorgen voor en lachen met jullie is fundamenteel aan mijn bestaan. Op momenten dat alles tegenzit (en die waren er zeker tijdens dit promotietraject) gaat er niets boven jullie kinderlijke onbezorgdheid, vrolijkheid en zelfs soms dwarsheid. Maret, "als ik jou niet had", zou dit proefschrift er niet zijn geweest. Dank voor het vertrouwen dat je me gaf toen ik na mijn bachelor natuurkunde en sterrenkunde nog wilde beginnen aan geneeskunde. Dank je me steunde toen ik (onbetaald) wetenschappelijke projecten wilde doen in plaats van werken voor geld. Dank voor je onophoudelijke steun toen promoveren erg veel uithouding van me vergde. Dank dat je bereid was om 3 maanden met mij naar New York te gaan en voor Vesper te zorgen terwijl je zwanger was van Doris, en voor de avonden en weekenden die ik soms moest werken terwijl jij de zorg voor onze kinderen op je nam. Met jou kan ik alles aan.

Curriculum Vitae

Wouter Anton Christiaan van Amsterdam was born on April 16, 1989 to Guus van Amsterdam and Ria van Amsterdam-Kloppenburg in Doetinchem (in ‘de Achterhoek’), the Netherlands. After finishing high school cum-laude, he moved to Utrecht to start an undergraduate degree in Physics & Astronomy at Utrecht University. Instead of moving on to a master degree in theoretical physics, Wouter decided to pursue a career in medicine, taking the long road of six more years at the University Medical Center Utrecht. During medical school, he participated in both the undergraduate and the graduate honours programmes, working on research projects first at the Anatomy department and later at the Radiology department. In the final year of medical school he visited the Artificial Intelligence in Medicine lab from Harvard Medical School for his graduation thesis. Winning the prestigious Alexandre Suerman stipend, Wouter followed his interest in applying machine learning for healthcare as a PhD-student at the radiology department of the University Medical Center Utrecht, advised by Prof. Dr. Pim de Jong, Prof. Dr. Tim Leiner, Dr. Joost Verhoeff and Rajesh Ranganath, PhD. During these 4 years, Wouter also graduated the post-graduate epidemiology master in Utrecht cum-laude, having chosen the medical statistics track. During his PhD Wouter was a visiting researcher at Rajesh Ranganath’s lab, part of CILVR, New York University. Wouter is a passionate researcher in the field of machine learning and causal inference for health care. In November 2021 he joined the Applied Sciences team of Babylon Health, London, as a senior research scientist.

Wouter is the recipient of several academic prizes, grants and awards, including the Utrecht University Bright mind award (2016), the thesis prize for epidemiology (“Frits de Waard penning”, 2020), NVIDIA-GPU seeding grants, the Alexandre Suerman stipend and a ZonMW grant for the PREMIUM project (co-applicant).

Wouter lives with his wife Maret and daughters Vesper and Doris in Odijk, the Netherlands.