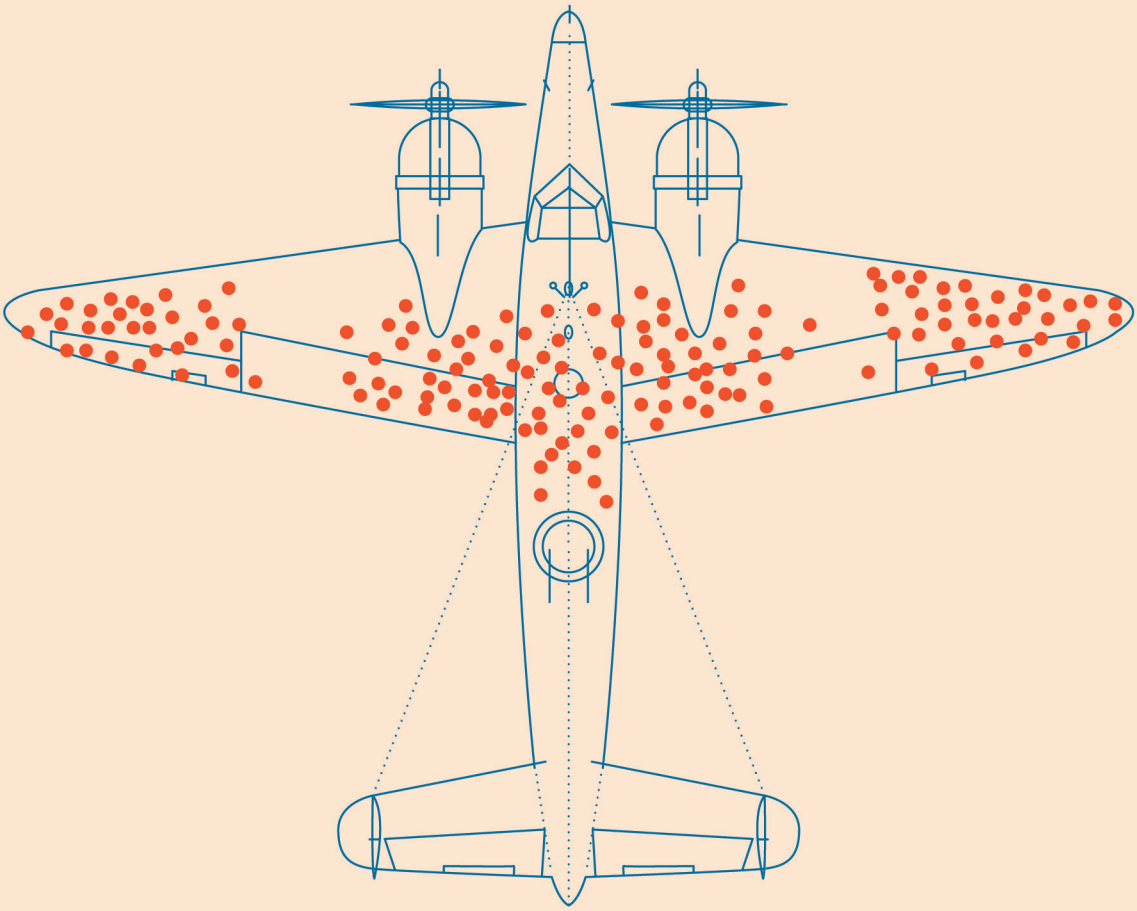
LEARNING HEALTHCARE SYSTEMS

Improving knowledge generation in healthcare



Wouter Bastiaan van Dijk

Learning healthcare systems

Improving knowledge generation in healthcare

Wouter Bastiaan van Dijk

Learning healthcare systems: Improving knowledge generation in healthcare

PhD thesis, Utrecht University, The Netherlands, with a summary in Dutch

Author:	Wouter Bastiaan van Dijk
Rights:	Please use the contents of this PhD thesis as much as possible (with appropriate references)
Layout and cover design:	Harma Makken, persoonlijkproefschrift.nl
Printing:	Ridderprint, www.ridderprint.nl
ISBN:	978-94-6416-786-3

This PhD thesis was printed with financial support of the Julius Center for Health Sciences and Primary Care.

Learning in healthcare systems

Improving knowledge generation in healthcare

Lerende zorgsystemen: het verbeteren van kennisgeneratie in de zorg

(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 24 februari 2022 des ochtends te 10.15 uur.

door

Wouter Bastiaan van Dijk

geboren op 27 december 1988
te Amsterdam

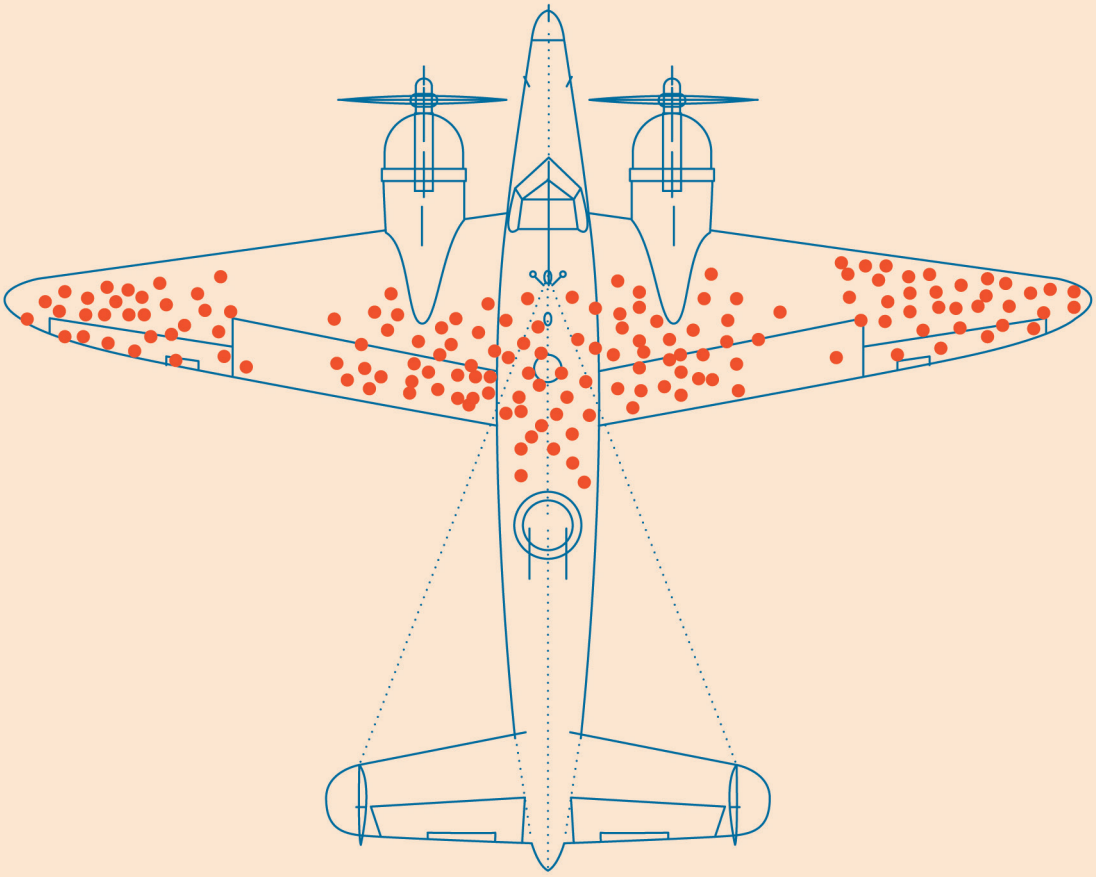
Promotoren:	Prof. dr. D.E. Grobbee
	Prof. dr. M.C. de Vries
Copromotoren:	Dr. E. Schuit
	Dr. R. van der Graaf

“The only true wisdom is in knowing you know nothing.”

- Socrates in Plato’s Apology

TABLE OF CONTENTS

Chapter 1.	General introduction & thesis outline	9
PART 1.	**CURRENT STATE OF CARDIOVASCULAR KNOWLEDGE**	**21**
Chapter 2.	Evidence substantiating the clinical practice guidelines of the European Society of Cardiology	23
Chapter 3.	Global applicability of the clinical practice guidelines of the European Society of Cardiology on general cardiology	57
PART 2.	**A BIRD'S EYE VIEW ON LEARNING HEALTHCARE SYSTEMS**	**73**
Chapter 4.	Learning Healthcare Systems: a scoping review of a methodological and ethical roadmap for implementation	75
Chapter 5.	The social value of knowledge in learning healthcare systems	93
Chapter 6.	The duty to support learning healthcare systems	109
PART 3.	**ROUTINELY COLLECTED DATA IN LEARNING HEALTHCARE SYSTEMS**	**117**
Chapter 7.	Text-mining electronic healthcare records for trial recruitment and baseline data collection	119
Chapter 8.	Text-mining electronic healthcare records for trial endpoint data collection	141
Chapter 9.	Dynamics in cardiac surgery: trends in population characteristics and the performance of the EuroSCORE II over time	159
Chapter 10.	A guide to privacy-preserving record linkage	177
Chapter 42.	General discussion, future perspectives & valorization	205
Addenda		**221**
	Summary	222
	Nederlandse samenvatting	225
	List of publications	229
	Dankwoord	230
	About the author	234



Chapter 1

General introduction & thesis outline

Over the last century, medical practice moved from solely being built on unsystematic, intuition based expert opinions to the current more scientifically substantiated form of evidence-based medicine (EBM), combining patient preferences, clinical circumstances, clinical expertise, and research evidence.[1,2] As a result of this shift, medicine is more grounded in scientific evidence than ever before.[1,3]

Still, as one of the primary pillars substantiating EBM, evidence generating research is often found to be slow, ill-aligned, and hard to translate to clinical practice.[4] Evidence is estimated to take an average of seventeen years before it finds itself implemented into care.[5] Moreover, from the large quantities of research published today only a fraction can be expected to translate to clinical practice.[5] For cardiology, the top five journals ranked by impact factor alone published 32,994 articles[1] over last five years (2016-2021; see **Supplemental materials, Table S1** for details), rendering implementation of all contributions highly unlikely.

During the COVID19 pandemic, the urgency for relevant knowledge has put a worldwide focus on generating evidence quickly. Studies that commonly took years to conduct were performed and implemented in clinical practice within a year.[6] This high speed and relevance of knowledge generation showed what can be achieved when research and clinical practice collaborate closely.[7] Still, of the clinical trials conducted that attempt to solve the recent COVID19 pandemic, less than one third was expected to influence clinical practice.[8] This is a surprising finding, as COVID19 is a new disease and every piece of research could thus have influenced clinical practice.[8]

Close collaboration between research and clinical practice is often considered to be the missing link, largely causing the slowness and limited relevance of research.[4,9] To mend this disconnect between research and clinical practice the Institute of Medicine (IoM) conceptualized learning healthcare systems (LHSs) in 2007.[4,7] In LHSs clinical practice and research are integrated to support each other in speeding up the generation of evidence for relevant clinical practice, and make implementation of this evidence in clinical practice easier.[10] This thesis focuses on the methods on which LHSs are based and will be based when brought into practice. It is therefore helpful to understand the quest for LHSs first. In the remainder of this introduction, I will describe how the use of knowledge and evidence in clinical practice developed over the last century. First, I will describe the paradigm of

1 A researcher or clinician would have to read an average of 30 articles per day to keep up with all publications from just these journals.

evidence-based medicine (EBM) more extensively. Second, I will explicate the concept of LHSs as introduced by the IoM. Finally, I provide the aim and an outline of this thesis.

EVIDENCE, MEDICINE, AND EVIDENCE-BASED MEDICINE

In the 20th century prosperity in the Western world led to an impressive number of milestone discoveries, making diseases previously considered untreatable now suddenly treatable.[11] Important discoveries included, among others, antibiotics, anticoagulants, blood types, electrocardiography, and even (coronary) arteriography.[12–14] Yet, newly discovered knowledge was only applied in clinical practice sparsely, and clinicians mainly delivered care based on their own experiences from their own patient care.[3,15] The latter meant that the evidence used to ground their medical practice in largely stemmed from anecdotes and case reports.[3] Observational studies, pioneered by researchers as John Snow and Florence Nightingale, were primarily used for population health research, and not performed in individual patient care.[16] Randomized controlled trials (RCTs) were relatively new still, with the first being conducted in 1948 on tuberculosis.[16]

The real impressive flight of scientific medicine came after the introduction of the principles of evidence-based medicine (EBM) in 1990. EBM was meant to bring structure to the use of evidence in clinical practice.[15] Evidence-based practice encouraged clinicians to use scientifically sound evidence as basis for their conduct in clinical practice. By educating clinicians in basic statistics and the trustworthiness of the various types of evidence, EBM was integrated into clinical practice swiftly.[3] Moreover, instruments like the evidence hierarchy were developed to help differentiate between the reliability levels of evidence, with RCTs and meta-analysis being considered the highest level of evidence and expert opinions the lowest level of evidence.

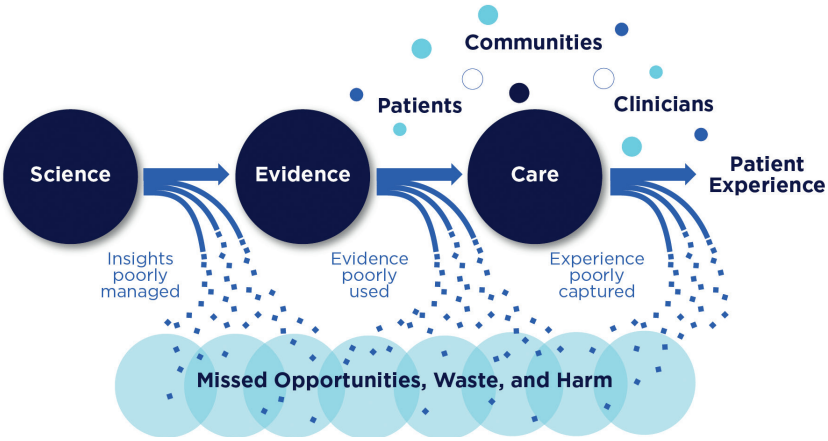
However, with the increased reliance on evidence also came the desire to speed up evidence generation and translation of evidence from research to clinical practice.[9] As mentioned before, evidence was found to take an average of seventeen years to be implemented in clinical practice after publication.[5] A delay reducing *state of the art* medicine into mere *near state-of-the-art-medicine*. Other trends like the possibility of using routinely collected (big) data instantly and a reduction in the abundance of research articles with limited relevance, made the call for a new paradigm louder at the early 2000s.[17]

Named as a possible cause for the slowness and misalignment of research and clinical practice was the absolute (legal and physical) split between research and clinical practice.[4] The separation between research and clinical practice inhibits desirable learning between the two. As a result, integration of research results in clinical practice is slow, and determining what research is most relevant for clinical practice is hard.[4] To mitigate these challenges the Institute of Medicine (IoM) conceived the concept of LHSs.[4]

FROM EBM TO LHSS

Main critique on EBM is that it is generally focused on linear knowledge generation and implementation, often leading to research waste (**Figure 1**).[9] Evidence is commonly generated in settings that are meant to reflect real clinical practice as closely as possible, however, these settings are often too strictly defined for the conduct of research.[18] Consequently, research populations often do not represent populations encountered in clinical practice with the result that findings from research have limited applicability in clinical practice.[18] Moreover, research is often conducted without direct input from clinical practice (i.e., doctors and especially patients) and does not always study the outcomes most relevant to patients or provide answers to the questions most relevant to clinical practice.[4,9]

Figure 1: Schematic of healthcare systems today[10]



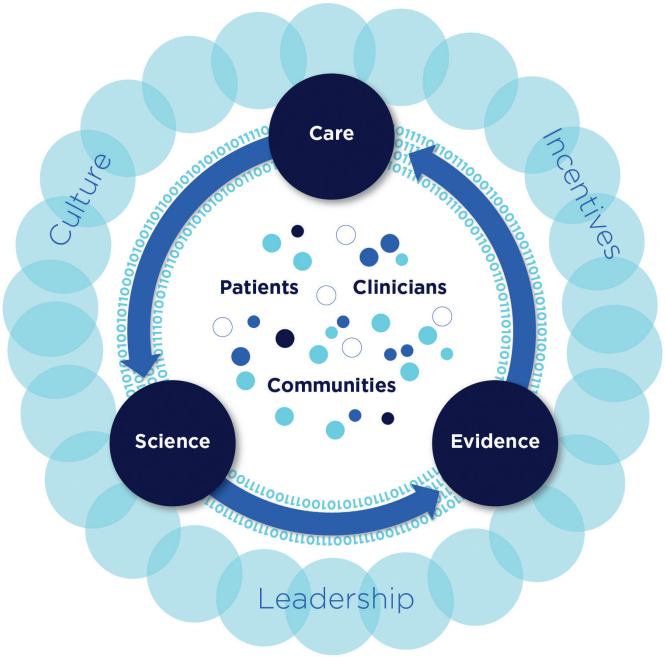
Printed with permission of The National Academies Press.

The IoM therefore introduced LHSs (**Figure 2**), meant to solve many of these challenges by integrating research and care in a systems approach through four characteristics: (1) using information technology (IT) to advance science, among others this includes using

routinely collected data, (2) creating patient-clinician partnerships, (3) setting the right incentives, and (4) creating a continuous learning culture.[10]

The four characteristics are meant to guide the transformation of current healthcare systems into LHSs.[10] By leveraging IT research clinical practice should be able to process larger amounts of data.[10] This includes the use of modern methods as artificial intelligence and machine learning algorithms.[19] Moreover, IT should allow research and clinical practice to simplify access to each other's data to allow learning from routinely collected data in during care delivery.[20] Patient-clinician partnerships are meant to help find the right questions to answer in research.[21] The right questions in this regard are questions found deemed relevant by patients and their clinicians, this includes choosing the right outcomes to test for in research.[22] Incentives should be set to stimulate research relevant for LHSs as a whole.[10] This would include moving away from rewarding individual researchers for their achievements, but instead focus on their contributions to the system at large.[21] Finally, a continuous learning culture should be created to allow learning on all levels in LHSs.[10] To achieve this, LHSs should support their patients, clinicians and researchers to explore new methods and train them in having the right mindset.[10]

Figure 2: Schematic of learning healthcare systems[10]



Printed with permission of The National Academies Press.

Since being introduced, much has been written about LHSs. To date, a PubMed search on learning healthcare systems delivers more than 60,000 results, with over 8,000 hits in 2020.[2] Yet, translation of LHSs literature is often experienced to be hard by researchers and policy makers alike.[23] In addition, few authors have focused on methods for LHSs and consequences they will have for knowledge generation.[23–25]

THESIS AIM AND OUTLINE

This thesis mainly focuses on cardiovascular care and the implications of introducing LHSs for this field. In particular, the aim is to examine methods of LHSs and their consequences for the knowledge generated in LHSs, resulting in the following research question:

"What are the consequences of using LHS methodologies for knowledge generation in the healthcare context?"

Cardiovascular care was chosen as field for this thesis as it has an extensive research body and is often thought to have a practice more grounded on EBM than other specialties. To extend our knowledge on the consequences of LHSs and methods to be applied in LHSs, I have conducted research on LHSs in three different directions.

Part 1 investigates the **current state of knowledge** in cardiovascular care. **Chapter 2** investigates the evidence base substantiating the clinical practice guidelines of the European Society of Cardiology (ESC). This chapter represents a baseline of the current cardiovascular evidence base in practice. **Chapter 3** investigates the applicability of ESC guidelines globally to assess how universally applicable knowledge is. Both chapters provide an overview of the evidence currently available in cardiovascular care and the extent to which evidence is applicable.

Part 2 investigates the **current state of implementation** of LHSs in practice and the consequences LHSs will have for research and clinical practice. **Chapter 4** investigates what is known about LHSs and implemented in practice already. **Chapter 5** investigates the consequences LHSs will have for knowledge generation, assessing whether faster and more local knowledge generation in LHSs changes the value of knowledge. **Chapter 6** investigates the duty of clinicians and researchers to participate in LHSs. A discussion which became suddenly topical when the COVID19 crisis asked for fast results and not all

2 Query: *learning AND health* AND system**, executed on August 9, 2021 latest.

clinicians and researchers readily aligned on reported results, consequently undermining research findings.

Part 3 investigates the **usage of routinely collected healthcare data** in LHSs. Most electronic healthcare records (EHR) data is stored in unstructured data formats. Usage of these data in research is sporadic still. I used text-mining methods to assess how applicable these data would be for clinical trials with these methods. **Chapter 7** investigates the applicability of text-mining methods on EHR data for trial recruitment and baseline data collection. **Chapter 8** investigates these same methods in EHR data for trial endpoint data collection. Next to using routinely collected healthcare data for trials, these data can also be used for prediction models. **Chapter 9** investigates how the performance of the EuroSCORE II risk score, a prediction of in-hospital death after cardiothoracic surgery, and patients' characteristics change over time. Finally, value can be added to datasets by linking them together. **Chapter 10** comprises an analysis of privacy-preserving record linkage methods. I assessed methods to perform this linkage in a privacy-preserving fashion.

REFERENCES

1. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. BMJ Evidence-Based Med. 2002;7(2):36-38.
2. Wieten S. Expertise in evidence-based medicine: a tale of three models. Philos Ethics, Humanit Med 2018 131. 2018;13(1):1-7.
3. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet. 2017;390(10092):415-423.
4. Institute of Medicine. The Learning Healthcare System. Washington, DC: The National Academies Press; 2007.
5. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med. 2011;104(12):510-520.
6. Agrawal G, Hermann R, Møller M, Poetes R, Steinmann M. Fast-Forward: Will the Speed of COVID-19 Vaccine Development Reset Industry Norms? New York: McKinsey & Company; 2021.
7. Romanelli RJ, Azar KMJ, Sudat S, Hung D, Frosch DL, Pressman AR. Learning Health System in Crisis: Lessons From the COVID-19 Pandemic. Mayo Clin Proc Innov Qual Outcomes. 2021;5(1):171-176.
8. Pundi K, Perino AC, Harrington RA, Krumholz HM, Turakhia MP. Characteristics and Strength of Evidence of COVID-19 Studies Registered on ClinicalTrials.gov. JAMA Intern Med. 2020;180(10):1398-1400.
9. Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. Lancet. 2014;383(9912):156-165.
10. Institute of Medicine. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. Washington, DC: The National Academies Press; 2013.
11. Harari YN. Sapiens: A Brief History of Humankind. New York: Harper; 2015.
12. Nabel EG, Braunwald E. A Tale of Coronary Artery Disease and Myocardial Infarction. N Engl J Med. 2012;366(1):54-63.
13. van Tellingen C. Chest pain and angina pectoris - or the ugly swan and the beautiful duckling. Netherlands Hear J. 2010;18(11):561-564.
14. Rhodes P, Thomson WAR, Underwood EA, Richardson RG, Guthrie DJ. History of medicine. In: Encyclopaedia Britannica. Chicago: Encyclopaedia Britannica; 2020.
15. Guyatt G, Cairns J, Churchill D, et al. Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine. JAMA. 1992;268(17):2420-2425.
16. Morabia A. History of epidemiological methods and concepts. In: Handbook of Epidemiology: Second Edition. Springer New York; 2014:43-74.
17. Big hopes for big data. Nat Med. 2020;26(1):1.
18. Gray E, Norris S, Schmitz S, O'Leary A. Do disparities between populations in randomized controlled trials and the real world lead to differences in outcomes? Futur Med. 2016;6(1):65-82.
19. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: Are we there yet? Heart. 2018;104:1156-1164.
20. van Staa TP, Klungel O, Smeeth L. Use of electronic healthcare records in large-scale simple randomized trials at the point of care for the documentation of value-based medicine. J Intern Med. 2014;275(6):562-569.
21. Maddox TM, Albert NM, Borden WB, et al. The Learning Healthcare System and Cardiovascular Care: A Scientific Statement from the American Heart Association. Circulation. 2017;134(14):e826–e857.
22. Harle CA, Lipori G, Hurley RW. Collecting, Integrating, and Disseminating Patient-Reported Outcomes for Research in a Learning Healthcare System. EGEMS (Washington, DC). 2016;4(1):1240.
23. Budrionis A, Bellika JG. The Learning Healthcare System: Where are we now? A systematic review. J Biomed Inform. 2016;64:87-92.

24. Abraham E, Blanco C, Lee CC, et al. Generating Knowledge from Best Care: Advancing the Continuously Learning Health System. NAM Perspectives. Discussion Paper. Washington, DC: National Academy of Medicine; 2016.

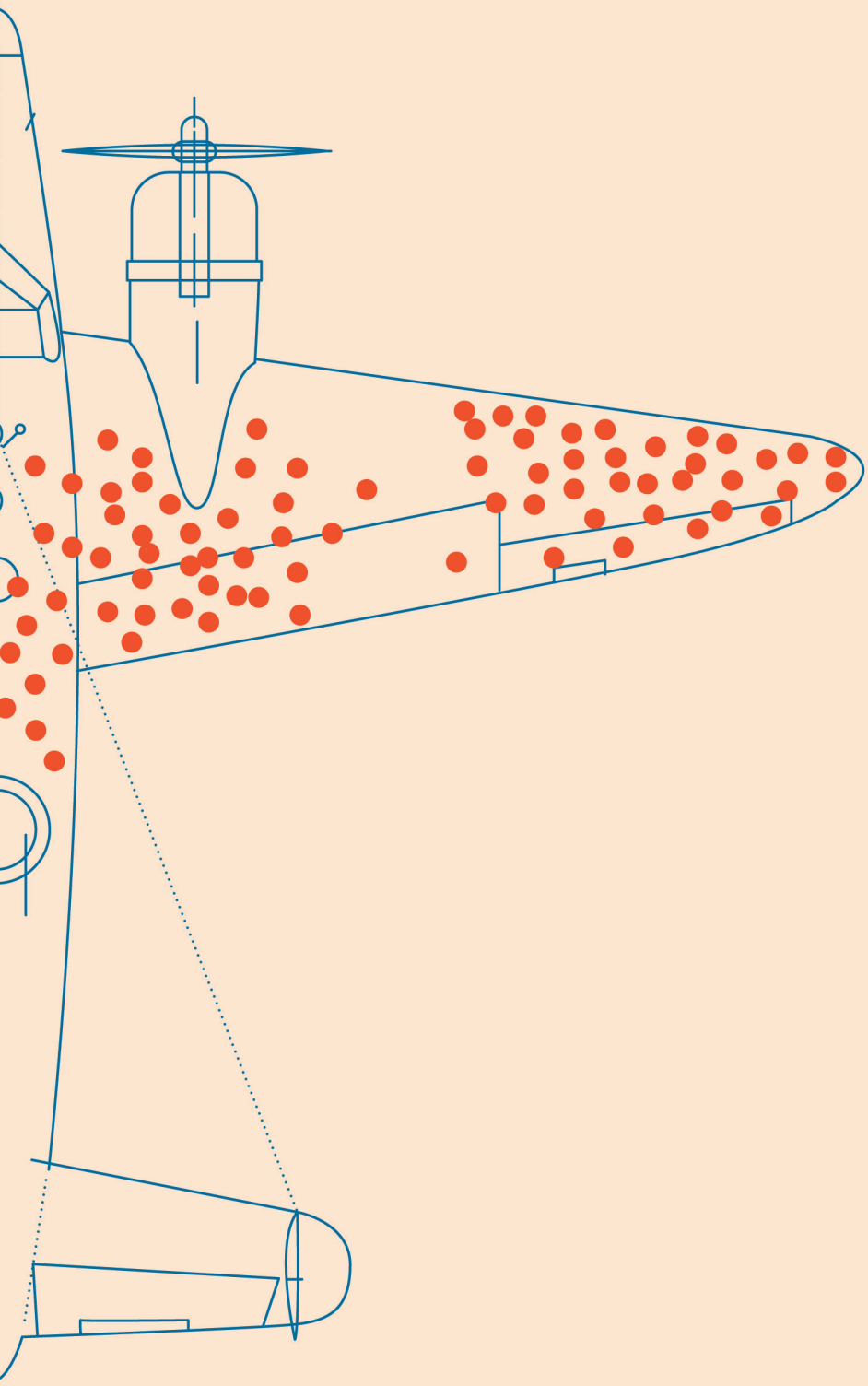
25. Krumholz HM. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. Health Aff. 2014;33(7):1163–1170.

SUPPLEMENTARY MATERIALS

Table S1: Number of publications for the top 5 highest impact factor journals in cardiology

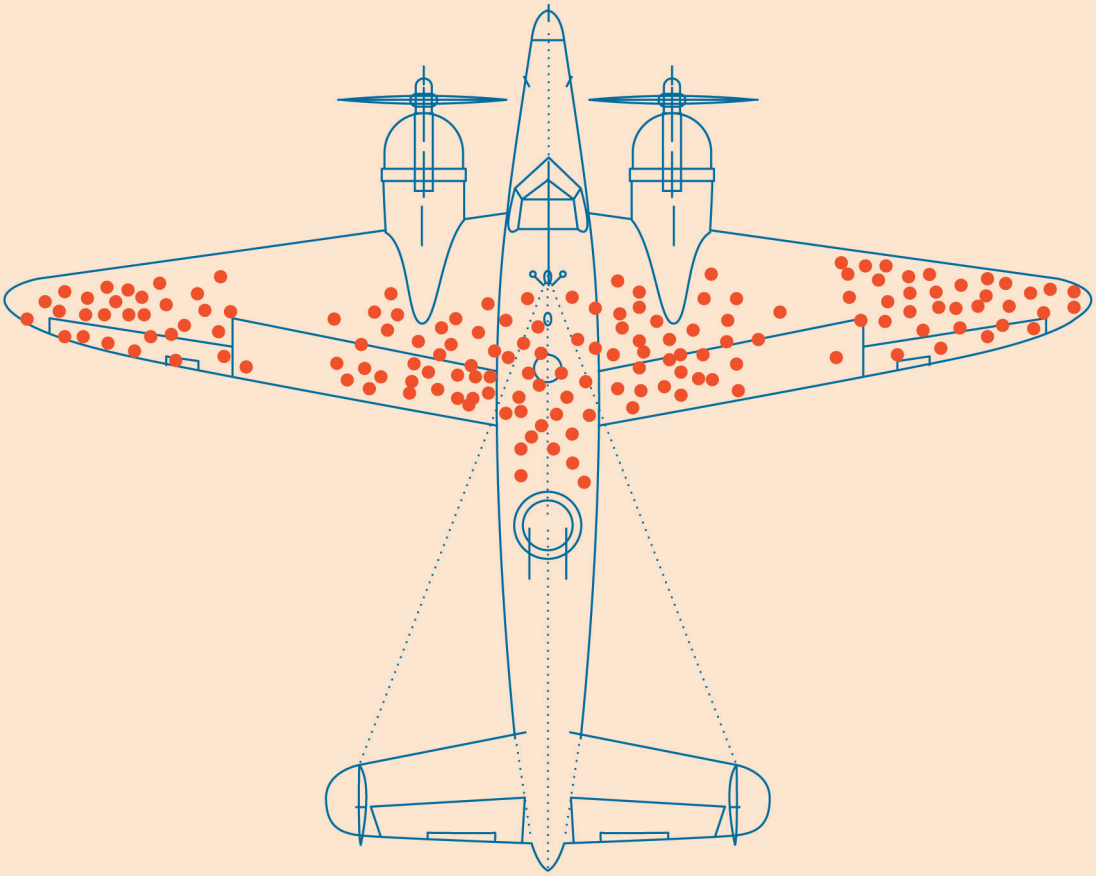
Journal	Impact factor	Publications (n)
Nature Reviews Cardiology	32.419	1,099
European Heart Journal	29.983	5,088
Circulation	29.690	10,392
Journal of the American College of Cardiology	24.094	14,180
Circulation research	17.367	2,235
Sum		32,994

Journals were selected on their Journal Impact Factor in the category Cardiac & Cardiovascular Systems – Scie on Clarivate Journal Citation Reports; number of publications retrieved from PubMed using journal titles plus [journal] as search string with time window set from July 1, 2016 to July 1, 2021. Analysis performed on July 7, 2021.



PART 1

Current state of cardiovascular knowledge



Chapter 2

Evidence substantiating the clinical practice guidelines of the European Society of Cardiology

Wouter B van Dijk, Diederick E Grobbee, Martine C de Vries,
Rolf H H Groenwold, Rieke van der Graaf, Ewoud Schuit

European Journal of Preventive Cardiology (2019); 26 (18): 1944–1952.

&

Wouter B van Dijk, Diederick E Grobbee

European Journal of Preventive Cardiology (2019); 26 (18): 1915–1917.
(see end of chapter)

ABSTRACT

Background

Reviews of clinical practice guidelines have repeatedly concluded that only a minority of guideline recommendations are supported by high-quality evidence from randomized controlled trials (RCTs). The aim is to evaluate if these findings apply to the whole cardiovascular evidence base or specific recommendation types and actions.

Methods

All recommendations from current European Society of Cardiology (ESC) guidelines were extracted with their class (I [treatment is beneficial], II [treatment possibly beneficial], III [treatment is harmful]) and Level of Evidence (LoE) (A [multiple RCTs/meta-analysis], B [single RCTs/large observational studies], C [expert opinion/small studies]). Recommendations were categorized by type (therapeutic, diagnostic, other) and actions (e.g. pharmaceutical intervention/non-invasive imaging/test).

Results

In total 3531 recommendations (median: 128 [IQR, 108-150]) were extracted from 27 guidelines. Therapeutic recommendations comprised 2545 (72.1%) recommendations, 411 (16.1%) were supported by LoE A, 833 (32.7%) by B and 1301 (51.1%) by C. Class I/III (should/should not) recommendations on minimal-invasive interventions were most supported by LoE A (55/183 [30.1%]; B: 70/183 [38.3%]; C: 58/183 [31.7%]), while class I/III recommendations on open surgical interventions were least supported by LoE A (15/164 [9.1%]; B: 34/164 [20.7%]; C: 115/164 [70.1%]). Of all (831 [23.5%]) diagnostic recommendations, just 44/503 (8.7%) class I/III recommendation were supported by LoE A (B: 125/503 [24.9%]; C: 334/503 [66.4%]).

Conclusions

Evidence levels supporting ESC guideline recommendations differ widely between recommendation types and actions. Attributing to this variability are different evidence requirements therapeutic/diagnostic recommendations, different feasibility levels for trials (e.g., open surgical/ pharmacological) and many off-topic/policy recommendations based on expert opinion.

INTRODUCTION

Clinical practice guidelines form the crest in translating science into clinical practice today. Clinicians report cardiovascular guidelines to be their main source of information for clinical decision making.[1] As such cardiovascular guidelines influence the care provided to millions of people worldwide.[2,3]

To justify this epistemological status, guidelines should be grounded in objective, high-quality evidence. Yet, a recent comparison of cardiovascular guidelines from the American College of Cardiology/American Heart Association (ACC/AHA) and European Society of Cardiology (ESC) published between 2008 and 2018 showed that a limited number of recommendations is supported by evidence from multiple high-quality Randomized Controlled Trials (RCTs; ACC/AHA: <10%, and ESC: <15%), and a majority by expert opinion and smaller studies.[4,5] These results fuel the criticism that guideline development lacks transparency on how recommendations are conceived[6–8], and previous claims that the evidence base underlying the cardiovascular guidelines is poor[5]. However, to know which paucities in the evidence base are problematic and where to focus improvement efforts to fill these, it is necessary to identify areas of recommendations not supported by high-quality evidence and identify the underlying reasons.

To reveal where gaps exist in the current cardiovascular evidence base, and allow better interpretation of the evidence underlying recommendations, this paper aims to identify which types of recommendations (e.g. therapeutic or diagnostic) and which recommended actions (e.g. pharmaceutical intervention or non-invasive imaging) are supported by which level of evidence (LoE) in the guidelines of the European Society of Cardiology (ESC).

METHODS

All documents referred to as Clinical Practice guideline were collected from the ESC website (https://www.escardio.org/Guidelines). Documents were categorized as comprehensive guidelines, focused updates, definition guidelines, position/expert consensus papers and other documents. Only current comprehensive guidelines and focused updates (short updates to comprehensive guidelines) were included for further analysis. Disease definitions and position papers were excluded from further analysis, since they were not considered representative for entire topics. The search and selection of guidelines was performed by one author (WvD).

From the remaining guideline documents, recommendations were extracted together with their corresponding class and evidence level (Box 1). Extraction was performed by one reviewer (WvD) using Tabula, version 1.2.1 (https://tabula.technology/)) and the results were exported to Excel, version 16.21.1. Only recommendations that contained a clearly indicated class and level of evidence were extracted. All recommendations were manually checked to have a recommendation text, class, and level of evidence. One author (WvD) categorized every guideline into a cardiovascular subspecialty area (electrophysiology, coronary artery disease, heart failure and myocardial disease, congenital and valvular heart disease, general cardiology, vascular medicine), and every recommendation by type (therapeutic, diagnostic, other [e.g. definitions]) and recommended action (main categories: pharmaceutical intervention, open surgical intervention, minimal-invasive/device intervention, non-invasive test/imaging, laboratory test, invasive test). Recommendations regarding preventive interventions were categorized as therapeutic and, similarly, recommendations about screening were regarded as diagnostic.

A random sample of recommendations was manually cross validated on extraction and classification by a second author (ES). Interrater reliability was assessed using Cohen's Kappa statistic and was considered sufficient (≥ 0.7), indicating further extraction could be performed by one author (WvD).

The overall number of recommendations, distributions and percentages of classes, evidence levels, types and recommended actions were calculated. Because of the high variety in recommendation numbers per guideline, these were summarized by calculating medians and interquartile ranges. For clarity recommendation types were also reported per cardiovascular subspecialty area and recommended action per type.

The ESC states to have published over 100 guidelines[9], while just 57 were published on the ESC website. We contacted the ESC to query about the remaining guidelines, whom pointed us to the website of the European Heart Journal (EHJ; https://academic.oup.com/eurheartj). A cross-reference of the EHJ website showed all current guidelines were published on the ESC website. We therefore focused only on guidelines as published on the ESC website.

Box 1: ESC classes of recommendations and levels of evidence

To appraise recommendations and value their underlying evidence the ESC uses a system of classes of recommendations and levels of evidence to grade.[10]

Recommendations are divided in the following classes:

I Evidence and/or general agreement that a treatment or procedure is beneficial, useful, effective.
Use of treatment or procedure is recommended.

II Conflicting evidence and/or a divergence of opinion about the usefulness/efficacy of a treatment or procedure.

a Weight of evidence/opinion is in favour of usefulness/efficacy. *Use of treatment or procedure should be considered.*

b Usefulness/efficacy is less well established by evidence/opinion. *Use of treatment or procedure may be considered.*

IIII Evidence or general agreement that a treatment or procedure is not useful/effective, and in some cases may be harmful.
Use of treatment or procedure is not recommended.

Evidence underlying recommendations is divided in the following levels of evidence:

A Data derived from multiple randomized clinical trials or meta-analyses.

B Data derived from a single randomized clinical trial or large non-randomized studies.

C Consensus of expert opinion and/or small studies, retrospective studies, registries.

RESULTS

On May 1st, 2019, 37 published documents were identified on the ESC website as current guidelines (Table 1 in the Supplementary Material). Ten documents were excluded because they concerned disease definitions (N=1), position/expert consensus papers (N=5), other documents (N=2) or did not contain clearly stated recommendations and evidence levels (N=2); leaving 27 guidelines for analysis.

The current 27 guidelines were published between 2003 and 2018 and provided 3531 recommendations. They comprised a median of 128 recommendations per guideline (IQR, 108-150) (Table 2 in the Supplementary Material), most recommendations were of class I (1684 recommendations; 47.7%), followed by class II (1577 recommendations; 44.7%) and III (270 recommendations; 7.6%). (**Figure 1**)

Of the 1,684 class I recommendations 360 (21.4%) recommendations were supported by LoE A, 489 (29.0%) recommendations by LoE B, and 835 (49.6%) recommendations by LoE C; of the 1577 class II 86 (5.5%) recommendations were supported by LoE A, 535 (33.9%) recommendations by LoE B, and 956 (60.6%) recommendations by LoE C; of the 270 class III 53 (19.6%) recommendations were supported by LoE A, 79 (29.3%) recommendations by LoE B, and 138 (51.1%) recommendations by LoE C.

Notably, the number of recommendations supported by LoE C varied widely between guidelines (range 36.7-76.4%).

Overall distributions of types and recommended actions

Overall, therapeutic recommendations comprised 2,545 recommendations (72.1%), diagnostic recommendations 831 recommendations (23.5%), and other recommendations 155 (4.4%) (**Figure 2**).

Among the 2545 therapeutic recommendations, 1134 (44.6%) recommendations were of class I, 1189 (46.7%) recommendations were of class II, and 222 (8.7%) recommendations were of class III. Of these, class I comprised 300 (26.5%) recommendations supported by LoE A, 350 (30.9%) recommendations supported by LoE B, and 484 (42.7%) recommendations supported by LoE C; class II comprised 63 (5.3%) recommendations supported by LoE A, 411 (34.6%) recommendations supported by LoE B, and 715 (60.1%) recommendations supported by LoE C; class III comprised 48 (21.6%) recommendations

supported by LoE A, 72 (32.4%) recommendations supported by LoE B, and 102 (45.9%) recommendations supported by LoE C.

Figure 1: Overall proportions of recommendation classes and levels of evidence

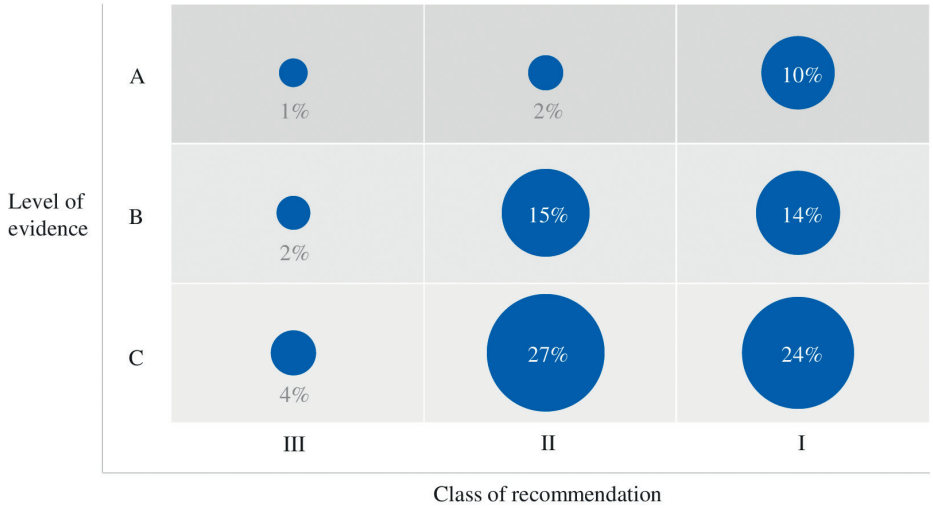
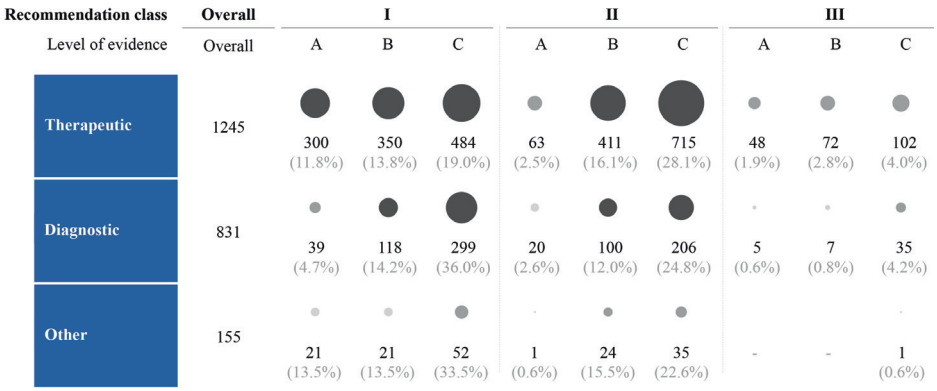


Figure 2: Proportions of types of recommendations by classes and evidence levels



The three therapeutic actions recommended most were pharmaceutical interventions (1245 recommendations, 48.9%), open surgical interventions (367 recommendations, 14.4%), and minimal invasive interventions (341 recommendations, 13.4%) (**Figure 3**). Pharmaceutical interventions comprised 236 (19.0%) recommendations supported by LoE A, 419 (33.7%) recommendations supported by LoE B, and 592 (47.6%) recommendations supported by LoE C; open surgical interventions comprised 20 (5.4%) recommendations

supported by LoE A, 78 (21.3%) recommendations supported by LoE B, and 269 (73.3%) recommendations supported by LoE C; comprised minimal-invasive interventions 63 (18.4%) recommendations supported by LoE A, 141 (41.3%) recommendations supported by LoE B, and 132 (40.2%) recommendations supported by LoE C.

Among the 831 diagnostic recommendations, 456 (54.9%) recommendations were of class I, 328 (39.5%) recommendations were of class II, 47 (5.7%) recommendations were of class III. Of these, class I comprised 39 (8.6%) recommendations supported by LoE A, 118 (25.9%) recommendations supported by LoE B, and 299 (65.6%) recommendations supported by LoE C; class II comprised 22 (6.7%) recommendations supported by LoE A, 100 (30.5%) recommendations supported by LoE B, and 206 (62.8%) recommendations supported by LoE C; class III comprised 5 (10.6%) recommendations supported by LoE A, 7 (14.9%) recommendations supported by LoE B, and 35 (74.5%) recommendations supported by LoE C.

The three diagnostic actions recommended most were non-invasive tests/imaging (378 recommendations, 45.6%), laboratory tests (156 recommendations, 18.8%), and invasive tests/imaging interventions (108 recommendations, 13.0%). Non-invasive tests/imaging comprised 14 (3.7%) recommendations supported by LoE A, 112 (29.6%) recommendations supported by LoE B, and 252 (66.7%) recommendations supported by LoE C; laboratory tests comprised 39 (25.0%) recommendations supported by LoE A, 32 (20.1%) recommendations supported by LoE B, and 85 (54.5%) recommendations supported by LoE C; invasive tests/imaging comprised 7 (6.5%) recommendations supported by LoE A, 19 (17.6%) recommendations supported by LoE B, and 82 (75.9%) recommendations supported by LoE C.

Distributions of types and recommended actions per subspecialty

General cardiology was the largest subspecialty with 875 (24.7%) recommendations, followed by coronary artery disease (726 recommendations, 20.6%) and Congenital and valvular heart disease (677 recommendations, 19.2%) (**Figure 4**). The largest proportion of class I recommendations supported by LoE A was found on coronary artery disease (169 recommendations, 23.3%). Congenital and valvular heart disease comprised the most recommendation supported by LoE C (505 recommendations, 74.6%) as result of a large number of recommendations on open surgical interventions.

Figure 3: Recommendations Modes of Action and Evidence Levels by Classes

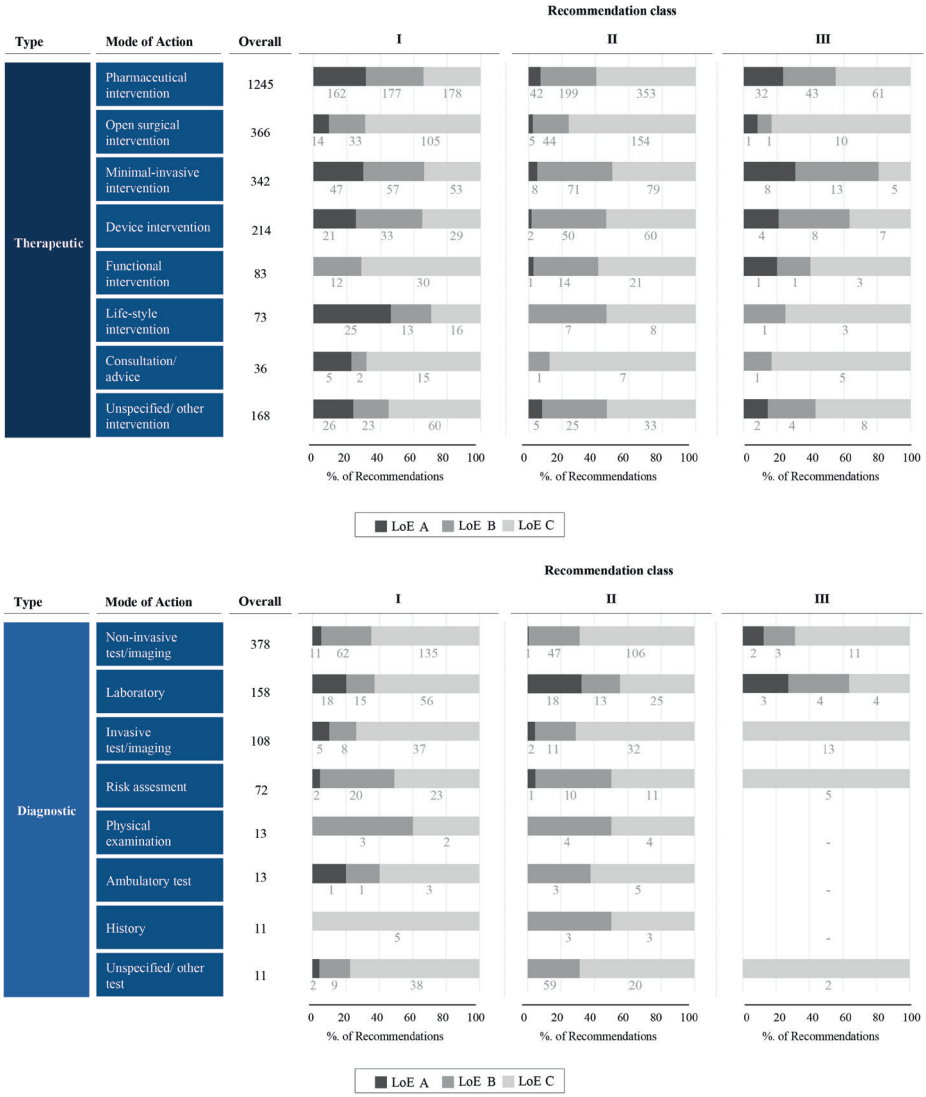
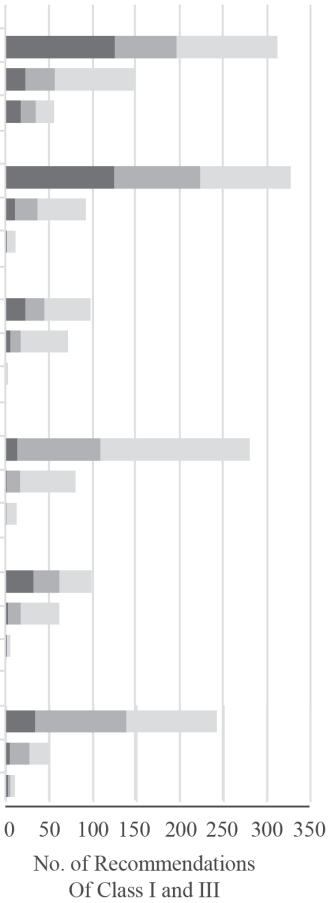


Figure 4: Recommendations by Subspecialty, Type and Level of Evidence

Recommednation type	Total No. of Recommendations
General Cardiology	
Therapeutic	312
Diagnostic	149
Other	55
Coronary artery disease	
Therapeutic	327
Diagnostic	92
Other	11
Vascular medicine	
Therapeutic	97
Diagnostic	71
Other	2
Congenital and valvular heart disease	
Therapeutic	280
Diagnostic	80
Other	12
Heart failure and myocardial disease	
Therapeutic	98
Diagnostic	61
Other	5
Electrophysiology	
Therapeutic	242
Diagnostic	50
Other	10



DISCUSSION

This in-depth analysis of the European Society of Cardiology guidelines shows that evidence levels supporting recommendations in cardiovascular guidelines vary widely per type, recommended action and subspecialty. Overall, just 14.1% of the recommendations are supported by multiple RCTs or meta-analysis (LoE A). However, when stratified to

their types and recommended actions we found that some recommendation groups are less substantiated by high-quality evidence than others. Therapeutic recommendations, in particular pharmaceutical, minimal-invasive and life-style recommendations, appear to be supported by higher quality evidence than diagnostic recommendations. We found recommendations on open surgical interventions, non-invasive tests/imaging and invasive tests/imaging to be least supported by high-quality evidence, attributing to the low evidence levels of recommendations in the ESC guidelines in general.

In their recent review Fanaroff et al. reported similar distributions of the overall LoEs in the cardiovascular guidelines of the ESC to those found in this review.[4] In addition, they reviewed the guidelines of the ACC/AHA of which they noted similar results. Yet the present study showed that these numbers and their accompanying conclusions do not apply to the evidence base as a whole. Instead, the quality of evidence supporting recommendations differs substantially between subspecialties, recommendation types and recommended actions. Systematic guideline analyses in other medical and surgical subspecialties have shown comparable distributions of few recommendations supported by level A evidence[11–17], and might, consequently, need comparable distinctions in the types and recommended actions.

A decade ago, Tricoci et al. already reported similar findings for the ACC/AHA guidelines. They identified several shortcomings in the organization of clinical research and guideline development process as possible explanations for this shortage in recommendations supported by high-quality evidence. These shortcomings included fragmentation of the research enterprise (a lack in common goals, vision and collaboration), missing incentives to fill evidence gaps and potential conflicts of interests. Fanaroff et al. found that the distributions of LoEs of the ACC/AHA guidelines did not change between 2008 and 2018, i.e. since the review by Tricoci et al., and that the ESC guidelines exhibited similar LoE distributions and trends over time.[5]

Implications for the cardiovascular evidence base

The findings of the present analysis provide focus for improvement efforts of the current cardiovascular evidence base.

Recommendations on pharmaceutical, minimal-invasive and life-style interventions were found to be most supported by high-quality evidence (30-40%) and put the low overall

evidence levels (15%) in a more positive perspective as such. Yet, it is still low when used to support adequate evidence-based decision making in practice and should be improved.

To improve the cardiovascular evidence base, more focus should be put on generating evidence for diagnostic recommendations and recommendations on open surgical interventions; two areas mainly supported by expert opinion and small studies still. Albeit lower evidence levels (LoE B) are understandable for research demonstrating diagnostic test accuracy, since these studies in general will have a cross-sectional design, the highest level (i.e. RCTs/meta-analysis) should be required to determine the consequences for patients of implementing a new diagnostic tool in clinical practice. Akin, it is important to also distinct guideline recommendations on their goal. Sometimes allowing lower evidence levels when only discussing test accuracies.

Hence, many recommendations on open surgical interventions lack support from high-quality evidence too. Evidence should be generated to fill these paucities to increase the reliability of recommendations on these interventions. Yielding research results to fill these evidence paucities on surgical treatments can be more difficult due methodological challenges. Justly executed surgical intervention trials are by nature more complex than pharmaceutical intervention trials due to the increased number of variables at play[18] and difficulties in blinding patients and doctors[19]. Undesirably, RCTs are relatively less common in surgery as a consequence.[20] Despite these challenges in performing surgical trials, they cannot condone in not evaluating surgical treatments because of the size of their on the lives of patients. Execution of surgical trials could for instance be improved by designing trials more pragmatically or to supplement them with the increasingly amounts of observational data available resulting from recent technological advancements, for instance by moving towards a Learning Healthcare Systems.[21]

Yet, it needs to be recognized that it might never be possible to support all guideline recommendations by the same (high) levels of evidence. For instance, recommendations to initiate treatment may need stronger evidence (e.g. RCTs and/or meta-analysis) than recommendations not to use specific treatments (e.g. a case series), since for the latter it will not always be feasible nor ethical to require these levels of evidence.

Implications for the development of guidelines

It is indisputable that we will always be in need for more high-quality evidence to fill paucities in the existing evidence base. If recommendations are important for clinical

practice they should be included, and efforts should be made to support it with evidence when evidence is lacking. Yet, besides improving the evidence base, guidelines should also be improved, handling the evidence paucities as good as possible. In over a decade time, distributions of LoEs in guidelines have barely changed.[4] This flat-line in LoEs might not only be maintained by the paucities in the cardiovascular evidence base, but could also be maintained as a consequence of the organization of the guideline development process. Guideline committees (i.e. Task Forces) should reflect on the contents of the recommendations they issue in their documents.

First, guideline committees should consider the large quantity of recommendations supported by LoE C that guidelines comprise. Currently, guideline authors have a so-called wide margin of appreciation giving them substantial freedom in the contents of recommendations they include in guidelines. Whether such a margin of appreciation should be allowed in guidelines is a matter of debate.[6,8] Regardless, this results in a wide range of covered topics in guidelines, inside and outside the field of cardiovascular care. Current guidelines for instance contain recommendations stating that the diagnosis and prognosis of a disease should always be explained to patients (class I, LoE C)[22]. General recommendations like these can be of importance to the homogenization process of different European practices, but are also at risk of stating the obvious, not rising above the level of presumed textbook knowledge. Guideline committees should be aware of these risks and consciously choose whether they want to issue such recommendations. Alike, deliberations should be made when issuing recommendations on adjacent specialties. Recommendations on vaginal delivery in healthy women (class I, LoE C)[23] and brain MRI when neurological examination indicates Parkinsonism, ataxia or cognitive impairment (class I, LoE C)[22] might give an impression of overestimation, compromising the trust in cardiovascular guidelines as a whole.

Second, guideline committees should reflect on the goal they have with their recommendations. One-third of the recommendations of the current guideline on Cardiovascular Disease Prevention in Clinical Practice comprises concern policy topics. These recommendations range from promoting heathy school diets (class I, LoE B) to increasing fuel taxes (class I, LoE C)[24]. Other recommendations cover measures to take against drink-driving (class I, LoE B)[24], advices against binge drinking (class I, LoE C)[10] and requirements of resident training programs (class I, LoE C)[25]. Although the social engagement showing from these recommendations is positive, they do not add value in direct patient care. Moreover, these recommendations often contain political opinions

nearly impossible to support with solid scientific evidence. The ESC describes guidelines as documents to help physicians weigh benefits and risks of diagnostic and therapeutic procedures[26] and since these recommendations comprise diagnostic nor therapeutic procedures, they might be better served in separate *policy guidelines*.

Additionally, it might be concluded that current LoEs are too crude, leaving uncertainties in the reliability of the underlying evidence. It is currently not possible to easily separate small trials, small and large observational studies from each other (both LoE B) and expert opinion from small studies (both LoE C). The initiative of the ACC/AHA to indicate the origin of evidence underlying evidence levels evidence (e.g., level C now indicates whether it is based on expert opinion or limited data)[27] should be more widely adopted to delineate the trustworthiness of LoEs supporting recommendations. Another strong initiative is the Grading of Recommendations Assessment, and Evaluation (GRADE) framework commonly used in Cochrane reviews to standardize guideline recommendations classes and LoEs. [28] The GRADE framework provides guidance on advising standard key factors for recommendations, their classes and LoEs to improve the quality, consistency and reliability of recommendations.

More in general, efforts should be made to increase transparency on the guideline development process and policies. In an attempt to compare the evidence used in cardiovascular guidelines to the evidence found by the systematic literature search done during the guideline development process we discovered that such a comparison was impossible since the search strategies used for evidence identification were not published by the ESC. Additionally, the governing policies noted that only peer-reviewed literature should be considered during the formal literature review.[26] Enforcing this policy, however, would conflict with the use of level C evidence in terms of expert opinion.

Study limitations

Several limitations should be held into account for the present study. First, we tried to select recommendation type categories which would speak for themselves and would thus be open for as little debate as possible. Yet, recommendations were still interpreted and categorized by hand, exposing the categorization process to possible misclassifications.

Second, the quality of evidence underlying the LoEs was not independently assessed in this study. It would therefore be possible that some LoEs falsely suggest lower qualities of evidence than actually used. For instance, evidence categorized as LoE B could consist

of (small) RCTs and of observational studies. Similarly, LoE C can contain expert opinion evidence and small studies as case series, etc. Alternatively, evidence standards could have been shifted in the years between different guidelines or interpreted differently by distinct guideline committees, consequently skewing the results of this study.

Third, this study only focused on the cardiovascular guidelines of the ESC. Although, the review by Fanaroff et al. showed similar distributions for ACC/AHA and ESC guidelines, one should be cautious in applying the salient findings of this study to studies on other medical and surgical societies. Nonetheless, the large category types and recommended actions, might give an indication for the structural distribution of evidence in the current knowledge base in general.

CONCLUSIONS

The evidence base underlying the cardiovascular guidelines of the ESC differs widely by recommendation types and recommended actions. Different reasons attribute to this high variability, including different evidence levels for therapeutic and diagnostic recommendations, different feasibility levels of trials for different interventions (e.g., open surgical vs. pharmacological interventions) and many off-topic/policy recommendations based on expert opinion. The cardiovascular research enterprise should focus on increasing evidence on diagnostic and open surgical topics by redesigning how evidence is generated, and by leveraging the increasingly amounts of data available, for instance in a Learning Healthcare System. Additionally, guideline authors should avoid issuing (off-topic) recommendations based on expert opinion/small studies as much as possible and clinical research should focus on incentivizing research on open surgical interventions.

REFERENCES

1. American College of Cardiology. Real-World Evidence : Perceptions of Cardiologists. *Cardiol Mag* 2018; 47:28-29.
2. Vaucher J, Marques-Vidal P, Waeber G, Vollenweider P. Population impact of the 2017 ACC/AHA guidelines compared with the 2013 ESH/ESC guidelines for hypertension management. *Eur J Prev Cardiol* 2018;25:1111–1113.
3. Pencina MJ, Navar-Boggan AM, D'Agostino RB, Williams K, Neely B, Sniderman AD, Peterson ED. Application of New Cholesterol Guidelines to a Population-Based Sample. *N Engl J Med* 2014;370:1422–1431.
4. Fanaroff AC, Califf RM, Windecker S, Smith Jr. S, Lopes, Renato D. Levels of Evidence Supporting American College of Cardiology/American Heart Association and European Society of Cardiology Guidelines, 2008-2018. *JAMA* 2019;27715:1069–1080.
5. Tricoci P, Allen JM, Kramer JM, Califf RM, Jr SCS. Scientific Evidence Underlying the ACC / AHA Clinical Practice Guidelines. *JAMA* 2009;301:831–841.
6. Bakris G, Sorrenntino M. Redefining Hypertension — Assessing the New Blood-Pressure Guidelines. *N Engl J Med* 2018;378:497–499.
7. Remaley AT, Shamburek RD, Amar M. The new ACC/AHA cardiovascular risk guidelines: Impact and controversies. *Clin Chem* 2014;60:1365–1367.
8. Shekelle PG. Clinical Practice Guidelines: What's Next? *JAMA* 2018;320:757–758.
9. Windecker S. The European Society of Cardiology Clinical Practice Guidelines. *Eur Heart J* 2017;38:2696–2697.
10. European Society of Cardiology, European Society of Hypertension. ESC/ESH Guidelines for the management of arterial hypertension. *Eur Heart J* 2018;39:3021–3104.
11. Alseiari M, Meyer KB, Wong JB. Evidence Underlying KDIGO (Kidney Disease: Improving Global Outcomes) Guideline Recommendations: A Systematic Review. *Am J Kidney Dis* Elsevier; 2016;67:417–422.
12. Feuerstein JD, Akbari M, Gifford AE, Hurley CM, Leffler DA, Sheth SG, Cheifetz AS. Systematic analysis underlying the quality of the scientific evidence and conflicts of interest in interventional medicine subspecialty guidelines. *Mayo Clin Proc* Elsevier; 2014;89:16–24.
13. Feuerstein JD, Akbari M, Gifford AE, Cullen G, Leffler DA, Sheth SG, Cheifetz AS. Systematic review: the quality of the scientific evidence and conflicts of interest in international inflammatory bowel disease practice guidelines. *Aliment Pharmacol Ther* John Wiley & Sons, Ltd (10.1111); 2013;37:937–946.
14. Brito JP, Domecq JP, Murad MH, Guyatt GH, Montori VM. The Endocrine Society Guidelines: When the Confidence Cart Goes Before the Evidence Horse. *J Clin Endocrinol Metab* Narnia; 2013;98:3246–3252.
15. Koh C, Zhao X, Samala N, Sakiani S, Liang TJ, Talwalkar JA. AASLD clinical practice guidelines: A critical review of scientific evidence and evolving recommendations. *Hepatology* John Wiley & Sons, Ltd; 2013;58:2142–2152.
16. Wright JD, Pawar N, Gonzalez JSR, Lewin SN, Burke WM, Simpson LL, Charles AS, D´Alton ME, Herzog TJ. Scientific Evidence Underlying the American College of Obstetricians and Gynecologists´ Practice Bulletins. *Obstet Gynecol* 2011;118:505–512.
17. Poonacha TK, Go RS. Level of scientific evidence underlying recommendations arising from the National Comprehensive Cancer Network clinical practice guidelines. *J Clin Oncol* American Society of Clinical Oncology; 2011;29:186–191.
18. Ergina PL, Cook JA, Blazeby JM, Boutron I, Clavien P-A, Reeves BC, Seiler CM. Challenges in evaluating surgical innovation Europe PMC Funders Group. *Lancet* 2009;374:1097–1104.
19. Wartolowska K, Judge A, Hopewell S, Collins GS, Dean BJF, Rombach I, Brindley D, Savulescu J, Beard DJ, Carr AJ. Use of placebo controls in the evaluation of surgery: Systematic review. *BMJ* 2014;348:1–15.

20. Gelijns AC, Ascheim DD, Parides MK, Kent KC, Moskowitz AJ. Randomized Trials in Surgery. *Surgery* 2009;145.

21. Ferguson RE, Fiore L. The Changing Face of Clinical Trials - Pragmatic trials. *N Engl J Med* 2016;375:454–463.

22. Brignole M, Moya A, Lange FJ de, Deharo J-C, Elliott PM, Fanciulli A, Fedorowski A, Furlan R, Kenny RA, Martín A, Probst V, Reed MJ, Rice CP, Sutton R, Ungar A, Dijk JG van, Torbicki A, Moreno J, Aboyans V, Agewall S, Asteggiano R, Blanc J-J, Bornstein N, Boveda S, Bueno H, Burri H, Coca A, Collet J-P, Costantino G, Díaz-Infante E, et al. 2018 ESC Guidelines for the diagnosis and management of syncope. *Eur Heart J* 2018;39:1883–1948.

23. Regitz-Zagrosek V, Roos-Hesselink JW, Bauersachs J, Blomström-Lundqvist C, Cífková R, Bonis M De, Iung B, Johnson MR, Kintscher U, Kranke P, Lang IM, Morais J, Pieper PG, Presbitero P, Price S, Rosano GMC, Seeland U, Simoncini T, Swan L, Warnes CA, Deaton C, Simpson IA, Aboyans V, Agewall S, Barbato E, Calda P, Coca A, Coman IM, Backer J De, Delgado V, et al. 2018 ESC Guidelines for the management of cardiovascular diseases during pregnancy. *Eur Heart J* 2018;39:3165–3241.

24. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, Cooney M-T, Corrà U, Cosyns B, Deaton C, Graham I, Hall MS, Hobbs FDR, Løchen M-L, Löllgen H, Marques-Vidal P, Perk J, Prescott E, Redon J, Richter DJ, Sattar N, Smulders Y, Tiberi M, Worp HB van der, Dis I van, Verschuren WMM. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J* 2016;37:2315–2381.

25. Neumann F-J, Sousa-Uva M, Ahlsson A, Alfonso F, Banning AP, Benedetto U, Byrne RA, Collet J-P, Falk V, Head SJ, Jüni P, Kastrati A, Koller A, Kristensen SD, Niebauer J, Richter DJ, Seferović PM, Sibbing D, Stefanini GG, Windecker S, Yadav R, Zembala MO, Wijns W, Glineur D, Aboyans V, Achenbach S, Agewall S, Andreotti F, Barbato E, Baumbach A, et al. 2018 ESC/EACTS Guidelines on myocardial revascularization. *Eur Heart J* 2018;40:87–165.

26. European Society of Cardiology. Governing Policies and Procedures for the Writing of ESC Clinical Practice Guidelines. 2017; Sophia, Antipolis, Biot, France.

27. Halperin JL, Levine GN, Al-Khatib SM, Birtcher KK, Bozkurt B, Brindis RG, Cigarroa JE, Curtis LH, Fleisher LA, Gentile F, Gidding S, Hlatky MA, Ikonomidis J, Joglar J, Pressler SJ, Wijeysundera DN. Further Evolution of the ACC/AHA Clinical Practice Guideline Recommendation Classification System: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* 2016;133:1426–1428.

28. Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, Schünemann HJ. GRADE: going from evidence to recommendations. *Br Med J* 2008;336:1049–1051.

SUPPLEMENTARY MATERIALS

Table S1: Documents published as current clinical practice guideline on the website of the European Society of Cardiology

Document (Year of publication)	**Type**	Status
Syncope (2018)[1]	Guideline	Included
ESC/ESH Arterial Hypertension (2018)[2]	Guideline	Included
Cardiovascular Diseases during Pregnancy (2018)[3]	Guideline	Included
ESC/EACTS Guidelines on Myocardial Revascularization (2018)[4]	Guideline	Included
Fourth Universal Definition of Myocardial Infarction(2018)[5]	Disease definition	Excluded
Acute Myocardial Infarction in patients presenting with ST-segment elevation (2017)[6]	Guideline	Included
Dual Antiplatelet Therapy (DAPT) (Focussed Update) (2017)[7]	Guideline	Included
Peripheral Arterial Diseases (2017)[8]	Guideline	Included
Valvular Heart Disease (2017)[9]	Guideline	Included
Dyslipidaemias (2016)[10]	Guideline	Included
CVD Prevention in Clinical Practice (2016)[11]	Guideline	Included
Acute and Chronic Heart Failure (2016)[12]	Guideline	Included
Atrial Fibrillation (2016)[13]	Guideline	Included
Position paper on cancer treatments & cardiovascular toxicity (2016)[14]	Position paper	Excluded
Acute Coronary Syndromes (ACS) in patients presenting without persistent ST-segment elevation (2015)[15]	Guideline	Included
Infective Endocarditis (2015)[16]	Guideline	Included
Pericardial Diseases (2015)[17]	Guideline	Included
Pulmonary Hypertension (2015)[18]	Guideline	Included
Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death (2015)[19]	Guideline	Included
ESC/ESA Guidelines on non-cardiac surgery: cardiovascular assessment and management (2014)[20]	Guideline	Included
Aortic Diseases (2014)[21]	Guideline	Included
Acute Pulmonary Embolism (2014)[22]	Guideline	Included
Hypertrophic Cardiomyopathy (2014)[23]	Guideline	Included
Diabetes, Pre-Diabetes and Cardiovascular Diseases developed with the EASD (2013)[24]	Guideline	Included
Stable Coronary Artery Disease (2013)[25]	Guideline	Included
Cardiac Pacing and Cardiac Resynchronization Therapy (2013)[26]	Guideline	Included
Grown-Up Congenital Heart Disease (2010)[27]	Guideline	Included
Device Therapy in Heart Failure (Focused Update) (2010)[28]	Guideline	Included
The Role of Endomyocardial Biopsy in the Management of Cardiovascular Disease (2007)[29]	Position paper	Excluded
Angiotensin Converting Enzyme Inhibitors in Cardiovascular Disease (Expert Consensus Document on) (2004)[30]	Position paper	Excluded
Antiplatelet Agents (Expert Consensus Document on the Use of) (2004)[31]	Position paper	Excluded

Table S1: *(continued)*

Document (Year of publication)	Type	Status
B-Adrenergic Receptor Blockers (Expert Consensus Document on) (2004)[32]	Position paper	Excluded
ACC/AHA/ESC Supraventricular Arrhythmias (2003)[33]	Guideline	Included
Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project (2003)[34]	Other	Excluded
Medical Practice Guidelines: Separating science from economics (2003)[35]	Other	Excluded
Chest Pain (Management of) (2002)[36]	Guideline (without structured recommendations)	Excluded
Neonatal Electrocardiogram (Guidelines for the interpretation of the) (2001)[37]	Guideline (without structured recommendations)	Excluded

Table S2: Classes and Levels of Evidence by Guideline and Type of Recommendation

		I			II			III		
Guideline (Year of publication)	**n**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
General cardiology	*875*	*146*	*103*	*203*	*27*	*138*	*194*	*18*	*19*	*27*
Syncope (2018)[1]	114	4	20	20		29	37	1	2	1
Therapeutic	53	2	8	5		17	18	1	1	1
Consultation	1						1			
Device intervention	27	2	4	2		7	10		1	1
Functional intervention	7					4	3			
Life-style intervention	1			1						
Minimal-invasive intervention	3		1				2			
Other/multiple intervention(s)	5		3	2						
Pharmaceutical intervention	9					6	2	1		
Diagnostic	57	2	12	11		12	19		1	
Ambulatory test	4					2	2			
History	4			2			2			
Invasive test	6	2				3	1			
Non-invasive test/imaging	26		6	7		2	10		1	
Other/multiple test(s)	1		1							
Physical examination	12		2	2		4	4			
Risk assessment	1					1				
Unspecified test	3		3							
Other	4			4						
Definition statement	1			1						
Differential diagnosis	3			3						
Arterial Hypertension (2018)[2]	135	41	16	29	1	20	15	7	2	4
Therapeutic	101	39	8	16	1	14	11	7	2	3
Device intervention	1								1	
Life-style intervention	12	6	2	2		1				1
Other/multiple intervention(s)	10	5		4			1			
Pharmaceutical intervention	63	22	6	9	1	7	9	6	1	2
Unspecified intervention	15	6		1		6	1	1		
Diagnostic	30	2	8	9		6	4			1

		I			II			III		
Guideline (Year of publication)	**n**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Laboratory test	4		2			1				1
Non-invasive test/imaging	20	2	3	6		5	4			
Risk assessment	5		3	2						
Unspecified test	1			1						
Other	4			4						
Definition statement	4			4						
Cardiovascular Diseases during Pregnancy (2018)[3]	148	2	8	71		2	53		1	11
Therapeutic	121	2	8	51		2	47		1	10
Consultation	12			3			3		1	5
Device intervention	2			1			1			
Functional intervention	14		1	6		2	5			
Life-style intervention	1						1			
Minimal-invasive intervention	10		1	2			7			
Open surgical intervention	13			5			8			
Other/multiple intervention(s)	2			2						
Pharmaceutical intervention	61	2	5	28			21			5
Unspecified intervention	6		1	4			1			
Diagnostic	26			20			6			
Invasive test	1			1						
Laboratory test	4			3			1			
Non-invasive test/imaging	12			8			4			
Risk assessment	7			7						
Unspecified test	2			1			1			
Other	1									1
Policy statement	1									1
Dyslipidaemias (2016)[10]	96	16	4	26	10	13	21	4	1	1
Therapeutic	50	9	4	4	2	12	13	4	1	1
Life-style intervention	2			1			1			
Pharmaceutical intervention	38	8	2	2	1	10	11	3		1
Unspecified intervention	10	1	2	1	1	2	1	1	1	
Diagnostic	43	6		21	8		8			
History	2			2						
Laboratory test	38	6		16	8		8			
Risk assessment	1			1						
Unspecified test	2			2						
Other	3	1		1		1				
Differential diagnosis	2	1		1						
Research results	1					1				
CVD Prevention (2016)[11]	178	41	34	19	15	34	26	3	4	2
Therapeutic	64	19	15	2	4	13	6	2	2	1
Consultation	3	3								
Functional intervention	2				1	1				
Pharmaceutical intervention	26	6	4		3	6	2	2	2	1
Unspecified intervention	10	1	5			3	1			
Life-style intervention	23	9	6	2		3	3			
Diagnostic	43	6	2	10	10	8	3	1	2	1

		I			II			III		
Guideline (Year of publication)	**n**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
History	2					1	1			
Laboratory test	23	6		6	9				2	
Non-invasive test/imaging	2					1		1		
Risk assessment	10		1	4	1	1	2			1
Unspecified test	6		1			5				
Other	71	16	17	7	1	13	17			
Policy statement	71	16	17	7	1	13	17			
Non-cardiac surgery (2014)[20]	121	7	11	32	1	27	29	1	6	7
Therapeutic	71	5	7	17	1	20	14		5	2
Device intervention	3			2						1
Functional intervention	2			2						
Life-style intervention	1			1						
Minimal-invasive intervention	7		2	1		2	1		1	
Open surgical intervention	10		1			5	4			
Other/multiple intervention(s)	6	1				4			1	
Pharmaceutical intervention	35	3	3	9	1	8	7		3	1
Unspecified intervention	7	1	1	2		1	2			
Diagnostic	43	2	4	11		6	13	1	1	5
Invasive test	7	1	2	2		1				1
Laboratory test	7					3	2	1		1
Non-invasive test/imaging	20	1		6		2	8		1	2
Risk assessment	2		2							
Unspecified test	7			3			3			1
Other	7			4		1	2			
Policy statement	7			4		1	2			
Diabetes and CVD (2013)[24]	83	35	10	6		13	13	2	3	1
Therapeutic	65	33	6	3		11	7	2	3	
Consultation	3		1	1		1				
Device intervention	1	1								
Life-style intervention	9	6	2						1	
Minimal-invasive intervention	5	2	1			2				
Open surgical intervention	1	1								
Pharmaceutical intervention	38	19	2			8	5	2	2	
Unspecified intervention	8	4		2			2			
Diagnostic	17	2	4	3		1	6			1
Laboratory test	6	1	3	2						
Risk assessment	7		1				5			1
Unspecified test	4	1		1		1	1			
Other	1					1				
Policy statement	1					1				
Coronary artery disease	*726*	*122*	*101*	*149*	*34*	*113*	*149*	*13*	*24*	*21*
Myocardial revasc (2018)[4]	238	40	30	40	16	45	51	4	7	5
Therapeutic	213	38	26	37	15	37	45	4	7	4
Consultation	2			2						
Device intervention	1						1			
Functional intervention	1			1						
Minimal-invasive intervention	65	16	9	11	3	7	10	3	5	1

		I			II			III		
Guideline (Year of publication)	**n**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Open surgical intervention	62	8	10	11	1	13	17			2
Other/multiple intervention(s)	4	2	1	1						
Pharmaceutical intervention	77	12	6	10	11	17	17	1	2	1
Unspecified intervention	1			1						
Diagnostic	25	2	4	3	1	8	6			1
Invasive test	4	1				3				
Laboratory test	2			1		1				
Non-invasive test/imaging	13	1	2	1	1	3	4			1
Risk assessment	6		2	1		1	2			
Acute STEMI (2017)[6]	160	26	19	47	8	15	29	3	9	4
Therapeutic	124	24	16	26	8	12	24	3	9	2
Functional intervention	5			3		1	1			
Life-style intervention	2	1		1						
Minimal-invasive intervention	32	10	6	4	1	1	6	2	2	
Open surgical intervention	1						1			
Other/multiple intervention(s)	5	1		1	1	1			1	
Pharmaceutical intervention	77	12	10	16	5	9	16	1	6	2
Unspecified intervention	2			1	1					
Diagnostic	29	1	3	16		3	4			2
Invasive test	6	1		4		1				
Laboratory test	4			4						
Non-invasive test/imaging	18		3	7		2	4			2
Other/multiple test(s)	1			1						
Other	7	1		5			1			
Policy statement	7	1		5			1			
DAPT (Focused Update) (2017)[7]	64	10	5	7	4	17	16	1	3	1
Therapeutic	62	10	5	6	4	17	16		3	1
Minimal-invasive intervention	1	1								
Other/multiple intervention(s)	1				1					
Pharmaceutical intervention	60	9	5	6	3	17	16		3	1
Diagnostic	1							1		
Laboratory test	1							1		
Other	1			1						
Policy statement	1			1						
ACS/NSTEMI (2015)[15]	134	30	23	26	6	16	25	1	4	3
Therapeutic	106	26	16	17	5	14	20	1	4	3
Device intervention	2	1					1			
Functional intervention	1			1						
Life-style intervention	1	1								
Minimal-invasive intervention	18	7	1	3	1	4	1		1	
Open surgical intervention	4	1	1	1		1				
Other/multiple intervention(s)	1				1					
Pharmaceutical intervention	77	15	14	12	3	8	18	1	3	3
Unspecified intervention	2	1				1				
Diagnostic	26	4	7	7	1	2	5			
Invasive test	5		3		1		1			
Laboratory test	6	1	2	2		1				

		I			II			III		
Guideline (Year of publication)	**n**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Non-invasive test/imaging	11	2	1	5			3			
Other/multiple test(s)	1	1								
Risk assessment	2		1			1				
Unspecified test	1						1			
Other	2			2						
Policy statement	2			2						
Stable CAD (2013)[25]	130	16	24	29		20	28	4	1	8
Therapeutic	58	15	12	6		13	6	3	1	2
Consultation	2			2						
Functional intervention	4					2	1	1		
Minimal-invasive intervention	21	7	5	1		5		1	1	1
Pharmaceutical intervention	29	7	7	2		6	5	1		1
Unspecified intervention	2	1		1						
Diagnostic	70	1	12	21		7	22	1		6
Ambulatory test	3			1			2			
Invasive test	8			3			4			1
Laboratory test	12	1	3	5			2	1		
Non-invasive test/imaging	44		7	11		7	14			5
Risk assessment	2		2							
Unspecified test	1			1						
Other	2			2						
Policy statement	2			2						
Vascular medicine	*318*	*21*	*30*	*99*	*4*	*63*	*81*	*6*	*4*	*10*
PAD (2017)[8]	128	13	10	30	4	33	28	3	2	5
Therapeutic	96	13	5	16	4	29	21	3	2	3
Life-style intervention	4	1	1	2						
Minimal-invasive intervention	27		1	4	2	13	5	1		1
Open surgical intervention	33	4		3	2	8	12	1	1	2
Pharmaceutical intervention	27	7	2	4		8	4	1	1	
Unspecified intervention	5	1	1	3						
Diagnostic	29		5	12		4	6			2
Invasive test	4			1			2			1
Laboratory test	1					1				
Non-invasive test/imaging	16		5	7		2	2			
Other/multiple test(s)	2			2						
Risk assessment	5			2		1	1			1
Unspecified test	1						1			
Other	3			2			1			
Differential diagnosis	1						1			
Policy statement	2			2						
Aortic Diseases (2014)[21]	119	4	5	57		13	36			4
Therapeutic	61	2	4	25		5	23			2
Life-style intervention	3		1				1			1
Minimal-invasive intervention	15	1		7		2	5			
Open surgical intervention	24		1	7		2	13			1
Other/multiple intervention(s)	2		1				1			
Pharmaceutical intervention	9			7		1	1			

		I			II			III		
Guideline (Year of publication)	**n**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Unspecified intervention	8	1	1	4			2			
Diagnostic	58	2	1	32		8	13			2
Laboratory test	5			1		2	1			1
Non-invasive test/imaging	32	1		20		3	8			
Risk assessment	7	1	1	2		1	1			1
Unspecified test	14			9		2	3			
Acute PE (2014)[22]	71	4	15	12		17	17	3	2	1
Therapeutic	38	2	10	7		9	7	2	1	
Minimal-invasive intervention	5					1	3	1		
Open surgical intervention	3			2			1			
Other/multiple intervention(s)	1					1				
Pharmaceutical intervention	27	2	10	4		7	2	1	1	
Unspecified intervention	2			1			1			
Diagnostic	33	2	5	5		8	10	1	1	1
Invasive test	1						1			
Laboratory test	5			2		1	1		1	
Non-invasive test/imaging	17	2	2	1		6	5	1		
Risk assessment	4		2	1		1				
Unspecified test	6		1	1			3			1
Cong. and valv. heart disease	*677*	*14*	*104*	*200*	*3*	*44*	*258*		*7*	*47*
Valvular Heart Disease (2017)[9]	111		11	33	2	8	53		2	2
Therapeutic	107		11	31	2	8	51		2	2
Minimal-invasive intervention	10		2	1			7			
Open surgical intervention	68		6	26	1	3	32			
Pharmaceutical intervention	24		2	3	1	5	10		2	1
Unspecified intervention	5		1	1			2			1
Diagnostic	4			2			2			
Invasive test	3			2			1			
Non-invasive test/imaging	1						1			
Infective Endocarditis (2015)[16]	128	1	50	21		13	30		1	12
Therapeutic	100		40	15		6	26		1	12
Device intervention	6		1	2			3			
Functional intervention	3						1			2
Minimal-invasive intervention	1		1							
Open surgical intervention	24		9	4		6	5			
Pharmaceutical intervention	66		29	9			17		1	10
Diagnostic	24	1	10	6		3	4			
Laboratory test	4	1		2			1			
Non-invasive test/imaging	20		10	4		3	3			
Other	4					4				
Policy statement	4					4				
Pericardial Diseases (2015)[17]	112	4	11	40	1	11	36		1	8
Therapeutic	79	4	8	19	1	9	30		1	7
Functional intervention	7			1			6			
Minimal-invasive intervention	7		1			4	2			
Open surgical intervention	17		2	7		3	5			
Other/multiple intervention(s)	6		2	3		1				

		I			II			III		
Guideline (Year of publication)	**n**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Pharmaceutical intervention	38	4	3	7	1	1	15		1	6
Unspecified intervention	4			1			2			1
Diagnostic	33		3	21		2	6			1
Invasive test	2			2						
Laboratory test	10		1	4		1	3			1
Non-invasive test/imaging	14		1	12			1			
Other/multiple test(s)	2			1		1				
Unspecified test	5		1	2			2			
Pulm. Hypertension (2015)[18]	221	9	28	61		12	91		3	17
Therapeutic	152	9	25	23		11	71		3	10
Consultation	2			2						
Functional intervention	2			1		1				
Life-style intervention	1									1
Minimal-invasive intervention	1			1						
Open surgical intervention	13			5			7			1
Other/multiple intervention(s)	2			1			1			
Pharmaceutical intervention	124	9	24	11		10	61		3	6
Unspecified intervention	7		1	2			2			2
Diagnostic	48		2	27		1	11			7
Invasive test	25			13		1	6			5
Laboratory test	2		1	1						
Non-invasive test/imaging	12		1	7			3			1
Other/multiple test(s)	1			1						
Risk assessment	1									1
Unspecified test	7			5			2			
Other	21		1	11			9			
Differential diagnosis	2						2			
Policy statement	19		1	11			7			
Grown-Up Cong. heart Disease (2010)[27]	105		4	45			48			8
Therapeutic	105		4	45			48			8
Minimal-invasive intervention	15			8			7			
Open surgical intervention	68		1	29			34			4
Pharmaceutical intervention	4		1				2			1
Unspecified intervention	18		2	8			5			3
HF and myocardial disease	*325*	*25*	*38*	*73*	*6*	*49*	*106*	*9*	*7*	*12*
Acute/Chron. HF (2016)[12]	185	23	15	41	6	36	39	9	6	10
Therapeutic	129	20	13	12	6	31	22	9	6	10
Consultation	2	2								
Device intervention	21	6	3			6	3	2		1
Functional intervention	5			1		1	1		1	1
Life-style intervention	1			1						
Minimal-invasive intervention	9	1	1	1	1	2	2		1	
Open surgical intervention	3			1			2			
Other/multiple intervention(s)	3			1		1	1			
Pharmaceutical intervention	81	11	9	5	5	21	11	7	4	8
Unspecified intervention	4			2			2			

		I			II			III		
Guideline (Year of publication)	**n**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Diagnostic	52	2	2	27		4	17			
Ambulatory test	1					1				
History	1					1				
Invasive test	11			3		1	7			
Laboratory test	5	1		4						
Non-invasive test/imaging	28	1	1	17		1	8			
Risk assessment	1		1							
Unspecified test	5			3			2			
Other	4	1		2		1				
Policy statement	4	1		2		1				
Hypertr. Cardiom. (2014)[23]	132		22	32		11	64		1	2
Therapeutic	66		9	14		6	35		1	1
Consultation	4		1	3						
Device intervention	11		2			3	5		1	
Functional intervention	2			1			1			
Life-style intervention	2			1			1			
Minimal-invasive intervention	5			1		1	3			
Open surgical intervention	8		1	2			5			
Pharmaceutical intervention	32		5	6		2	18			1
Unspecified intervention	2						2			
Diagnostic	63		13	16		5	28			1
Ambulatory test	3			2			1			
Invasive test	11		1	3			6			1
Laboratory test	9		3	1			5			
Non-invasive test/imaging	34		6	9		4	15			
Other/multiple test(s)	2		1				1			
Risk assessment	4		2	1		1				
Other	3			2			1			
Policy statement	3			2			1			
Device Therapy in Heart Failure (Focused Update) (2010)[28]	8	2	1			2	3			
Therapeutic	8	2	1			2	3			
Device intervention	8	2	1			2	3			
Electrophysiology	*610*	*32*	*113*	*111*	*12*	*128*	*168*	*7*	*18*	*21*
Atrial Fibrillation (2016)[13]	151	14	23	10	5	49	38	3	5	4
Therapeutic	122	11	17	7	4	39	33	3	4	4
Consultation	3			1			2			
Functional intervention	6		3	2			1			
Life-style intervention	2	1					1			
Minimal-invasive intervention	18	1	1	1		9	6			
Open surgical intervention	10				1	3	6			
Pharmaceutical intervention	77	9	12	3	3	24	15	3	4	4
Unspecified intervention	6		1			3	2			
Diagnostic	26	2	6	3	1	9	4		1	
History	1					1				
Laboratory test	5	1			1	2			1	
Non-invasive test/imaging	14		5	1		4	4			

		I			II			III		
Guideline (Year of publication)	**n**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Physical examination	1		1							
Risk assessment	5	1		2		2				
Other	3	1				1	1			
Policy statement	3	1				1	1			
Ventr. Arrhythmias and Sudden Cardiac Death (2015)[19]	246	11	48	65	1	42	65	3	5	6
Therapeutic	182	8	30	44	1	37	53	3	4	2
Consultation	2			1			1			
Device intervention	65	5	12	11	1	15	20	1		
Functional intervention	9		1	6		1	1			
Life-style intervention	5		1	4						
Minimal-invasive intervention	43	1	11	5		14	9		2	1
Open surgical intervention	4		1	1			2			
Other/multiple intervention(s)	2					1	1			
Pharmaceutical intervention	50	2	4	14		6	19	2	2	1
Unspecified intervention	2			2						
Diagnostic	50	2	15	16		3	9		1	4
Ambulatory test	2	1	1							
History	1			1						
Invasive test	13		2	3		1	3			4
Laboratory test	3			2			1			
Non-invasive test/imaging	24	1	9	6		2	5		1	
Other/multiple test(s)	1			1						
Risk assessment	2		2							
Unspecified test	4		1	3						
Other	14	1	3	5		2	3			
Differential diagnosis	1					1				
Policy statement	12	1	3	5		1	2			
Research results	1						1			
Cardiac Pacing and Cardiac Resync. Therapy (2013)[26]	66	3	10	11	2	17	13	1	5	4
Therapeutic	65	3	10	11	1	17	13	1	5	4
Device intervention	65	3	10	11	1	17	13	1	5	4
Diagnostic	1				1					
Invasive test	1				1					
Supraventricular Arrhythmias (2003)[33]	147	4	32	25	4	20	52		3	7
Therapeutic	147	4	32	25	4	20	52		3	7
Device intervention	1	1								
Functional intervention	13		7	5		1				
Life-style intervention	3					3				
Minimal-invasive intervention	23		13	2		4	3			1
Open surgical intervention	1			1						
Other/multiple intervention(s)	4			4						
Pharmaceutical intervention	96	3	12	9	4	12	49		2	5
Unspecified intervention	6			4					1	1

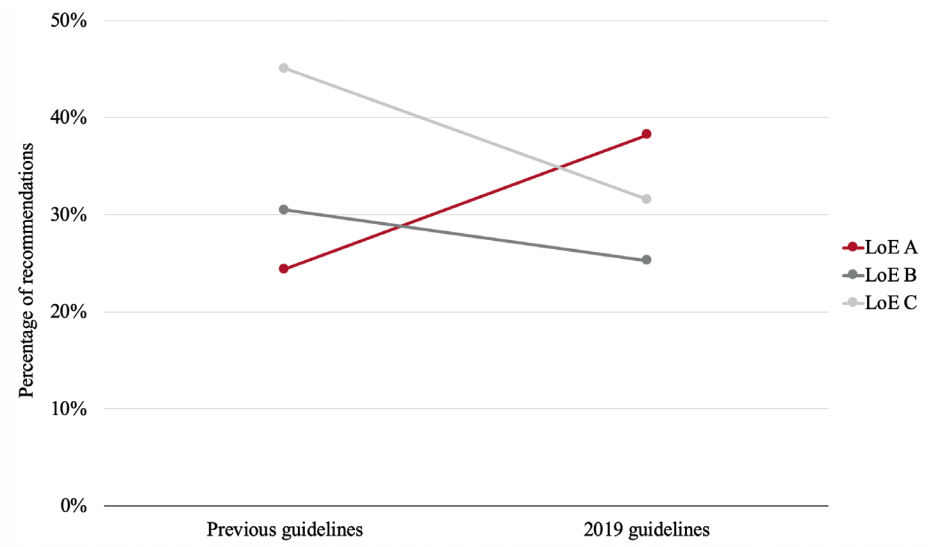
EDITORIAL

CARDIOVASCULAR GUIDELINES BASED ON HIGH-QUALITY EVIDENCE: ARE WE GETTING THERE?

Less than 15% of all recommendations in cardiovascular guidelines are supported by high-quality evidence from multiple randomized clinical trials (RCTs) and meta-analyses, i.e., level A evidence.[1–3] One recent study showed that the evidence levels underlying the European and American guidelines did not change over the last decade.[2] Considering that similar evidence levels have been found in most medical and surgical subspecialties a grim outlook on the medical evidence base as a whole arises. However, light comes after the darkest hour.

Evidence levels of the recently released 2019 guidelines of the European Society of Cardiology (ESC) prudently set forth a more commendable trend. In the five new 2019 ESC guidelines the number of *should* and *should not* recommendations (class I and III) supported by level of evidence (LoE) A increases from 24.4% (IQR 18.9%-38.5%) to 38.2% (IQR 22.8%-45.9%), while reducing the number of recommendation supported by LoE B and C (**Figure 1**).[4–8] Simultaneously the number of recommendations was raised from a median of 96 (Interquartile range [IQR] 83-130) in previous guidelines to a median of 120 (IQR 120-138) recommendations in the 2019 guidelines.

Figure 1: Overall evidence levels supporting should and should not (class I and III) recommendations



These results suggest that the efforts of the last decade to improve the evidence on the management of cardiovascular disease might start to bear fruit. Indeed, these numbers represent just one year of guideline releases of one major cardiovascular society and should thus be interpreted with care. Nonetheless, it should not be unrecorded that steps were taken in the right direction by the authors of the 2019 guidelines. To further understand and improve the cardiovascular evidence base and identify where gaps exist, guidelines should be broken-down on their actions recommended, for example by categorizing recommendations by intervention (pharmaceutical/open surgical/life-style/etc) and diagnostic (laboratory test/invasive imaging/risk stratification/etc). Findings from such a recent break-down analysis of the ESC guidelines showed that more than two-thirds of the 3,531 recommendations issued between 2003 and 2018 were on therapeutic topics (largest groups: pharmaceutical [48.9%], open surgical [14.4%] and minimal-invasive [13.4%] interventions) and one-third on diagnostic topics (largest groups: non-invasive tests/imaging [45.5%], laboratory tests[18.9%] and invasive tests/imaging [13.0%] interventions). Pharmaceutical and lifestyle interventions were expectedly substantially more grounded in level A evidence (class I/III respectively 15.6% and 34.3%) than open surgical (class I/III 4.1%) and diagnostic recommendations (range class I/III 2.3-13.4%).[3]

Ideally, all care delivered would be supported by evidence from well conducted RCTs. It should thus not be accepted to leave high-quality evidence levels around 40%. However, RCT's may not be the best methodology to, for example, demonstrate the value of diagnostic or prognostic assessments.[9] Trials are inevitable when diagnostic and prognostic research induce different intervention strategies or when adequate reference standards are missing.[10] In reclassification, accuracy and cost-effectiveness studies however, trials offer no surplus over the results of cross-sectional and therapeutic studies combined often making observational research the highest level of evidence viable for diagnostic and prognostic recommendations.[10] Hence, rendering the lower evidence levels attached to these studies controversial.

Trials also have remained laborious and become increasingly expensive, barely innovating in their execution, limiting their numbers and leavening clinicians and guideline authors to rely on lower levels of evidence as a result.[11] Innovation in trials to decrease their costs is sorely needed to allow them to add knowledge on evidence gaps where profitable business cases lack. Since most costs of trials are found to lie within their (manual) patient accrual and follow-up large opportunities to decrease these costs can be found in creating learning healthcare systems by leveraging routinely collected data as found in electronic

healthcare records.[12] Until then guideline authors are forces to decide on whether or not they issue recommendations based on weaker evidence instead, keeping guidelines patient-centric and avoiding them to become a mere recital of facts.[13]

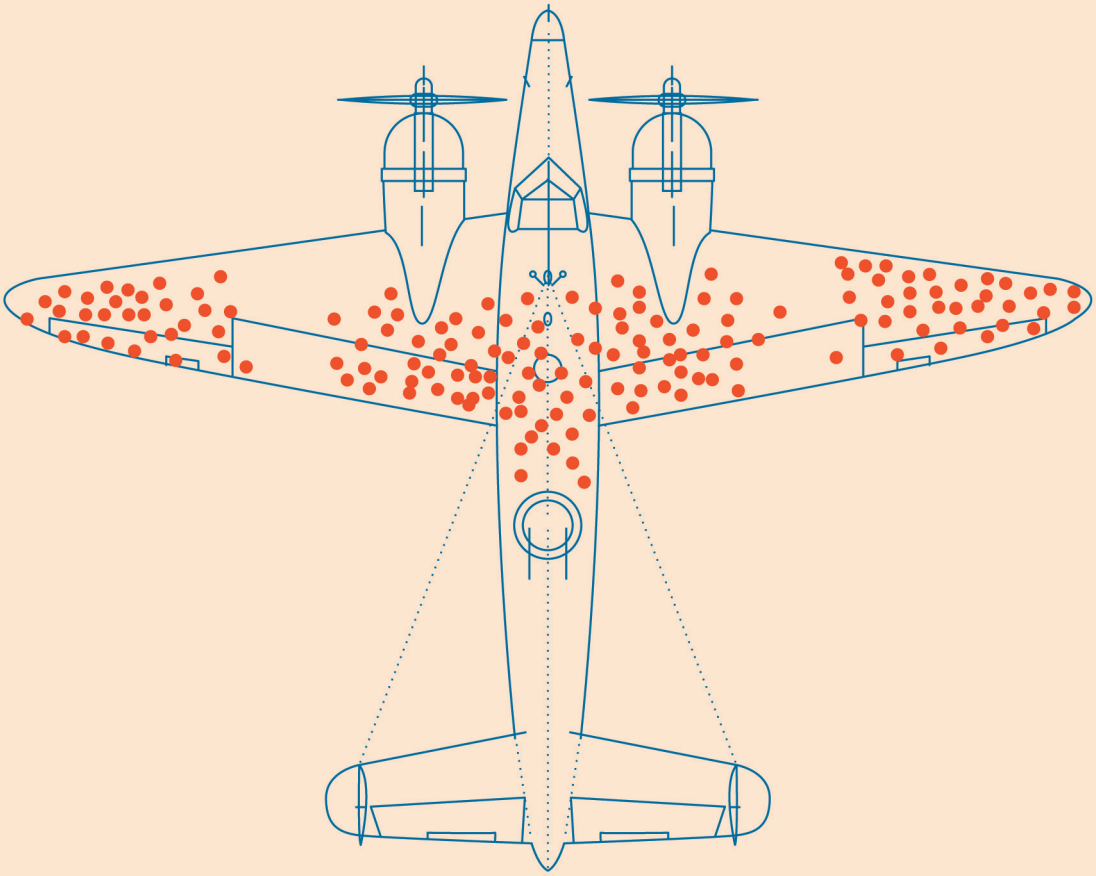
By issuing recommendations substantiated by weaker evidence however, guideline authors do not get the privilege to release just any recommendation. While a margin of appreciation for guideline authors remains necessary as in some cases a recommendation just supported by trusted colleagues (LoE C) is better than no recommendation, guideline recommendations should remain focused on their goal to support clinicians in their practice. Recommendations on for instance political topics should therefore be considered carefully before issuing.

Evidence levels supporting guidelines give insights on the status of the cardiovascular evidence base at a given time. To grow these insights, guidelines should be broken-down on their actions (e.g. pharmaceutical interventions, surgical interventions, diagnostic imaging, prognostic stratification) to identify where to focus further efforts on improving the cardiovascular evidence base. Previously, pharmaceutical and lifestyle interventions were for example found to be expectedly more grounded in level A evidence than diagnostic recommendations.[3] Yet, these new guidelines represent the current status of guideline development and unburdened from their predecessors history they suggest that improvements in guideline development are on their way.

REFERENCES

1. Tricoci P, Allen JM, Kramer JM, Califf RM, Jr SCS. Scientific Evidence Underlying the ACC / AHA Clinical Practice Guidelines. JAMA 2009;301:831–841.
2. Fanaroff AC, Califf RM, Windecker S, Smith Jr. S, Lopes, Renato D. Levels of Evidence Supporting American College of Cardiology/American Heart Association and European Society of Cardiology Guidelines, 2008-2018. JAMA 2019;27715:1069–1080.
3. Dijk WB van, Grobbee DE, Vries MC de, Groenwold RHH, Graaf R van der, Schuit E. A systematic breakdown of the levels of evidence supporting the European Society of Cardiology guidelines. Eur J Prev Cardiol 2019; 26(18):1944-52.
4. Cosentino F, Grant PJ, Aboyans V, Bailey CJ, Ceriello A, Delgado V, Federici M, Filippatos G, Grobbee DE, Hansen TB, Huikuri H V, Johansson I, Jüni P, Lettino M, Marx N, Mellbin LG, Östgren CJ, Rocca B, Roffi M, Sattar N, Seferović PM, Sousa-Uva M, Valensi P, Wheeler DC, Piepoli MF, Birkeland KI, Adamopoulos S, Ajjan R, Avogaro A, Baigent C, et al. 2019 ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD. Eur Heart J 2019; 41(2):255-323.
5. Knuuti J, Wijns W, Saraste A, Capodanno D, Barbato E, Funck-Brentano C, Prescott E, Storey RF, Deaton C, Cuisset T, Agewall S, Dickstein K, Edvardsen T, Escaned J, Gersh BJ, Svitil P, Gilard M, Hasdai D, Hatala R, Mahfoud F, Masip J, Muneretto C, Valgimigli M, Achenbach S, Bax JJ, Neumann F-J, Sechtem U, Banning AP, Bonaros N, Bueno H, et al. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. Eur Heart J 2019; 41(2): 407-77.
6. Brugada J, Katritsis DG, Arbelo E, Arribas F, Bax JJ, Blomström-Lundqvist C, Calkins H, Corrado D, Deftereos SG, Diller G-P, Gomez-Doblas JJ, Gorenek B, Grace A, Ho SY, Kaski J-C, Kuck K-H, Lambiase PD, Sacher F, Sarquella-Brugada G, Suwalski P, Zaza A, Potter T De, Sticherling C, Aboyans V, Basso C, Bocchiardo M, Budts W, Delgado V, Dobrev D, Fitzsimons D, et al. 2019 ESC Guidelines for the management of patients with supraventricular tachycardiaThe Task Force for the management of patients with supraventricular tachycardia of the European Society of Cardiology (ESC). Eur Heart J 2019; 41(2):655-702.
7. Konstantinides S V, Meyer G, Becattini C, Bueno H, Geersing G-J, Harjola V-P, Huisman M V, Humbert M, Jennings CS, Jiménez D, Kucher N, Lang IM, Lankeit M, Lorusso R, Mazzolai L, Meneveau N, Ní Áinle F, Prandoni P, Pruszczyk P, Righini M, Torbicki A, Belle E Van, Zamorano JL, Galié N, Gibbs JSR, Aboyans V, Ageno W, Agewall S, Almeida AG, Andreotti F, et al. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS). Eur Heart J 2019; 41(4):543-603.
8. Mach F, Baigent C, Catapano AL, Koskinas KC, Casula M, Badimon L, Chapman MJ, Backer GG De, Delgado V, Ference BA, Graham IM, Halliday A, Landmesser U, Mihaylova B, Pedersen TR, Riccardi G, Richter DJ, Sabatine MS, Taskinen M-R, Tokgozoglu L, Wiklund O, Mueller C, Drexel H, Aboyans V, Corsini A, Doehner W, Farnier M, Gigante B, Kayikcioglu M, Krstacic G, et al. 2019 . Eur Heart J 2019; 41(1):111-88.
9. Grobbee DiE, Hoes AW. Clinical epidemiology: Principles, methods, and applications for clinical research. Sudbury, Mass: Jones and Bartlett Publishers; 2014.
10. Distraction From Randomization in Diagnostic Research. Ann Epidemiol Elsevier; 2006;16:540–544.

11. Moore TJ, Zhang H, Anderson G, Alexander GC. Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015-2016. JAMA Intern Med American Medical Association; 2018;178:1451.
12. Bentley C, Cressman S, Hoek K van der, Arts K, Dancey J, Peacock S. Conducting clinical trials—costs, impacts, and the value of clinical trials networks: A scoping review. Clin Trials SAGE PublicationsSage UK: London, England; 2019;16:183–193.
13. Piepoli MF, Ceconi C. Levels of evidence in the European Society of Cardiology Guidelines: Gaps in knowledge? Eur J Prev Cardiol SAGE PublicationsSage UK: London, England; 2019;1941-3.



Chapter 3

Global applicability of the clinical practice guidelines of the European Society of Cardiology on general cardiology

Wouter B van Dijk, Ewoud Schuit, Rieke van der Graaf, Rolf H H Groenwold, Sara Laurijssen, Barbara Cassadei, Marco Roffi, Seye Abimbola, Martine C de Vries, Diederick E Grobbee

Submitted.

ABSTRACT

Background

Clinical practice guidelines of the European Society of Cardiology (ESC) are widely endorsed. However, the applicability of recommended actions may vary between countries. We investigated the applicability of ESC guidelines on general cardiology areas by surveying local feasibility to comply with the recommended actions in 102 countries and assessed how the applicability differed as a function of countries income levels.

Methods

All recommendations from seven ESC guidelines on general cardiology areas were extracted and labeled on recommended actions. A survey was sent to all 102 ESC national cardiac societies (NCSs). Respondents were asked to score recommended actions on their availability in clinical practice on a four-point Likert scale (fully available, mostly/often available, mostly/often unavailable, fully unavailable), and select the top three barriers potentially impairing national availability. Applicability was assessed overall, per World Bank gross national income (GNI) level, and per guideline.

Results

A total of 875 guideline recommendations on general cardiology was extracted. Responses were received from 64 of 102 (62·7%) ESC NCSs. On average, 77·4% (95%-confidence interval [CI] 34·6-100·0) of the actions were fully available, 11·4% (95%-CI 0·0-28·0) mostly/often available, 6·2% (95%-CI 0·0-22·9) mostly/often unavailable, and 5·0% (95%-CI 0·0-23·4) fully unavailable. In low-income countries (LICs) substantially more actions were fully unavailable (29·4% [95%-CI 0·0-100·0]) compared to HICs (2·4% [95%-CI 0·0-15·4]). Notably, actions with lowest availability scores were often fully or mostly unavailable independent of GNIs. ESC guidelines on general cardiology comprised more actions fully unavailable in LICs (22·4% [95%-CI 0·0-50·3]) compared with high-income countries (HICs; 1·4% [95%-CI 0·0-33·4]). Actions were most often not available due to financial barriers.

Conclusions

Availability of actions recommended and applicability of ESC guidelines on general cardiology is high in HICs and low in LICs, being inversely correlated with country gross national incomes.

INTRODUCTION

Morbidity and mortality related to cardiovascular disease (CVD) have decreased substantially in the past half-century.[1] Nonetheless, CVD remains one of the leading causes of death globally.[1] Countries with lower gross national incomes (GNI) per capita have higher morbidity and mortality rates for CVD with broad differences between countries.[1]

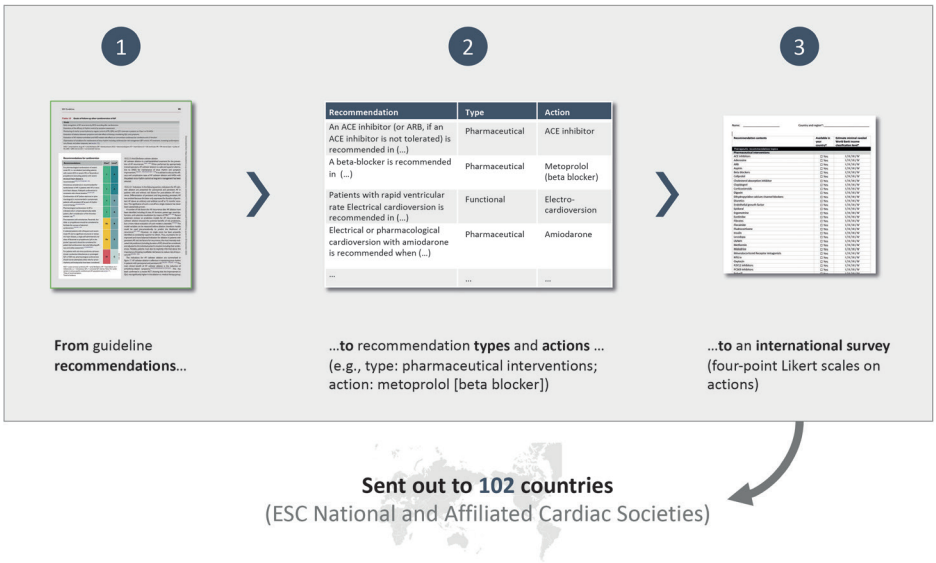
To improve CVD outcomes, clinical practice guidelines encourage clinicians to provide care according to evidence-based standards.[2,3] In optimizing global cardiovascular care, guidelines of major international cardiovascular societies play an important role.[4,5] The European Society of Cardiology (ESC) aspires to support cardiovascular care globally, as reflected with her 102 member and affiliated national cardiac societies (NCSs), representing more than 80% of the world population.[6] Comprehensive implementation of ESC guidelines would have considerable impact on the burden of cardiovascular disease worldwide. However, for guidelines to be applicable globally, recommendations in these guidelines need to be available across countries.

To assess applicability of the current ESC guidelines recommendations in general cardiology areas, we surveyed the 102 ESC NCSs and investigated whether availability of recommended actions was influenced by country income levels.

METHODS

We extracted all guideline recommendations from ESC clinical practice guidelines on general cardiology areas (Diabetes, Pre-Diabetes and CVD [2013]; Non-Cardiac Surgery [2014]; CVD Prevention [2016]; Dyslipidemias [2016]; CVD during pregnancy [2018]; Arterial Hypertension [2018]; Syncope [2018]) and categorized them into actions (e.g., “prescribe metoprolol”), and associated recommendation types (e.g., pharmaceutical intervention). Guidelines were defined as being on general cardiology according to classifications used previously.[7,8] Only guidelines on general cardiology were included as these were considered to be most relevant in a global context. Subsequently, we disseminated a survey to all 102 ESC NCSs to score recommended action on their availability (**Figure 1**).

Figure 1: Data collection methods



Collecting guideline recommendations

Guideline recommendations, actions and associated type groups were collected in a similar fashion as described previously.[7,8]

In short, current Clinical Practice Guidelines were downloaded from the ESC website (https://www.escardio.org/Guidelines). All documents were categorized as comprehensive practice guidelines, focused updates, definition guidelines, position papers and other documents by one author (WD). Guidelines were categorized to their cardiovascular subspecialty using the same classification as previously by these and other authors.[7,8] Guidelines on general cardiology were included for further analysis. From every guideline, all recommendation texts, classes and Levels of Evidence (LoEs) were retrieved using Tabula (version 1.2.1, https://tabula.technology) by one author (WD). The results were exported to Microsoft Excel (version 16.21.1.1) and manually checked on having a recommendation text, class and LoE.

Collecting guideline recommendation actions

To be of use to clinical practice, guideline recommendations need to be actionable (e.g., recommend subscribing a beta blocker or a primary percutaneous intervention [PCI]), to which each recommendation can be labeled (e.g., “prescribe metoprolol” or

"perform PCI"). For the current study, we derived actions from recommendation texts, and associated type groups (e.g., pharmaceutical intervention or minimal-invasive intervention). For example, when the name of a substance was available this was taken instead of its substance group (e.g., "metoprolol" instead of "beta blocker"; the entire list of actions can be found in **Table S1 of the supplementary materials**).

After extraction and categorization, a random sample of actions and types was manually cross validated by a second author (ES). Interrater reliability was assessed using Cohen's Kappa statistic and was considered sufficient (≥ 0·7).

Survey on availability of recommended ESC guideline actions

A survey comprising a list of all recommended actions was sent out to all 102 ESC NCSs to be completed by local country experts on cardiovascular disease (the full survey can be found in **the supplementary materials**). Survey responses were collected between June 2020 and June 2021. NCSs that did not respond to the survey received monthly reminders and were actively contacted by e-mail and phone by members of the ESC Global Affairs committee.

To assess recommendation applicability, NCSs were asked to score every recommended action on its availability in their country on a four-point Likert scale (fully available [score of 4], mostly/often available [3], mostly/often unavailable [2], fully unavailable [1]). In addition, NCSs were asked for each recommendation type group (e.g., pharmaceutical intervention or laboratory tests) to indicate the top three reasons for their lack of availability. Reasons for actions being unavailable were derived from consensus discussions within the author group and the ESC Global Affairs office.

Analyzing guideline applicability

The overall number of action availability scores was calculated and aggregated by response options (i.e., score of 1 to 4). When multiple responses were received from one NCS, averages of these responses were taken.

Responses were reported by World Bank income level (high income [HIC]: >$12,055 per capita; upper-middle income [UMIC]]: $3,895-12,055 per capita; lower-middle income [UMIC]: $996-3,895 GNI per capita; low-income [LIC]: <$996 GNI per capita; reference year: 2019)[9], country and guideline.

To assess clinical guideline applicability, actions were re-mapped to their original recommendations and aggregated on their corresponding guideline.

Reasons for actions not being available were reported by recommendations type group.

RESULTS

Guideline recommendations and actions

A total of 34 documents were retrieved from the ESC website, on May 1, 2020 latest. Of these, 27 were categorized as comprehensive guidelines, comprising 3,351 recommendations. Seven guidelines with 875 (26·1%) recommendations were identified as current guideline on general cardiology areas. These 875 recommendations were used for further analysis.

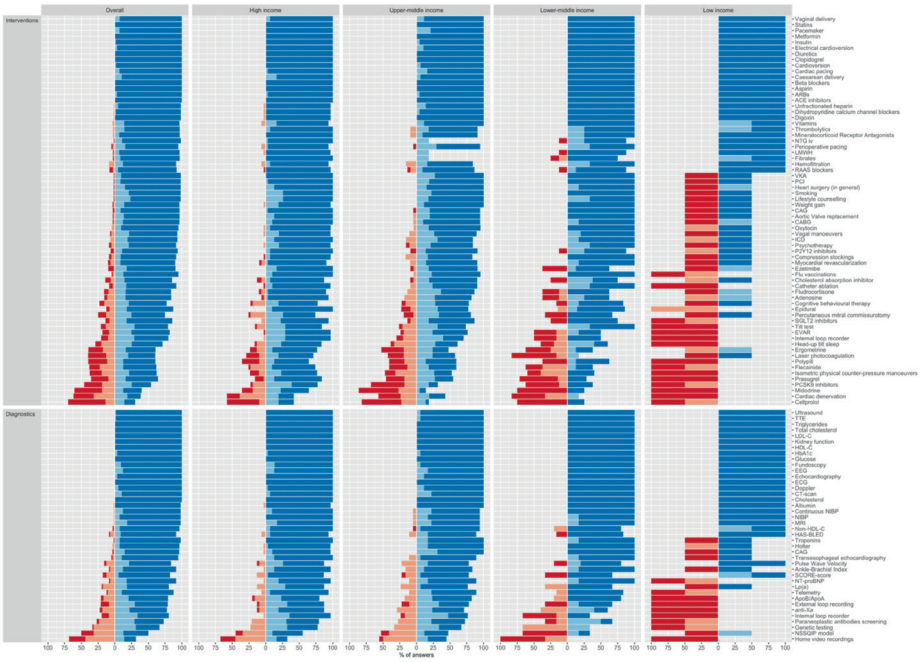
From the 875 recommendations, 139 different actions were extracted. Most actions were on therapeutic pharmaceutical interventions (n=38; 27%), diagnostic non-invasive tests (n=18; 13%) and laboratory tests (n=18; 13%). For 91 recommendations it was not possible to retrieve specific actions; these recommendations comprised policy statements (n=80), disease definitions (n=5), differential diagnosis (n=5) and research results (n=1).

Action availability

Responses were received from 64 (HIC:32, UMIC: 22, LMIC: 8, LIC: 2) out of 102 countries that were contacted.

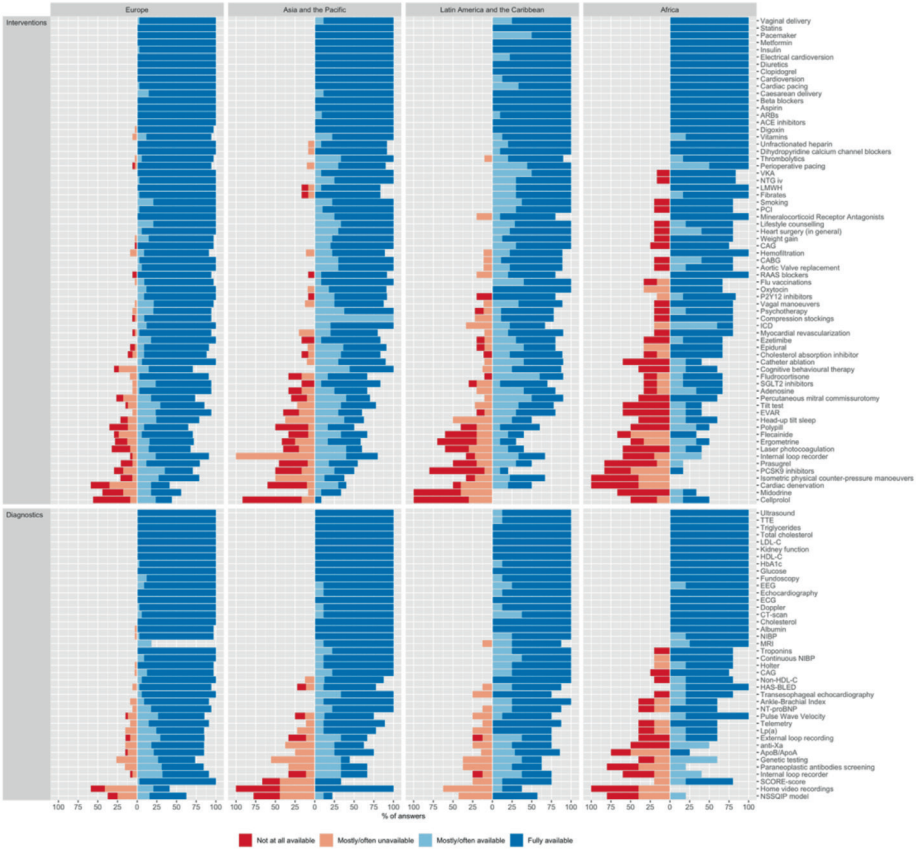
Overall, 77·4% (95%-confidence interval [CI] 34·6-100·0) of the actions were fully available, 11·4% (95%-CI 0·0-28·0) mostly/often available, 6·2% (95%-CI 0·0-22·9) mostly/often unavailable, and 5·0% (95%-CI 0·0-23·4) fully unavailable (**Figure 2**). Of recommended actions on interventions 69·1% (95%-CI 6·7-100·0) were fully available, 10·7% (95%-CI 0·0-36·6) mostly/often available, 6·6% (95%-CI 0·0-32·7) mostly/often unavailable, and 13·6% (95%-CI 0·0-62·2) fully unavailable. Of recommended actions on diagnostics 75·9% (95%-CI 15·7-100·0) were fully available, 8·5% (95%-CI 0·0-32·9) mostly/often available, 6·9% (95%-CI 0·0-33·8) mostly/often unavailable, and 8·7% (95%-CI 0·0-52·3) fully unavailable.

Figure 2: Action availability per type and country income level



Availability of recommended actions decreased substantially with World Bank income levels (for action availability per country see **Figure S1 of the supplementary materials**), in high income countries 2·4% (95%-CI 0·0-15·4) of the actions was scored fully unavailable, in upper-middle income countries 4.7% (95%-CI 0·0-26·7), in lower-middle income countries 10·6% (95%-CI 0·0-48·2), and in low-income countries 29·4% (95%-CI 0·0-100·0). Notably, actions with lowest availability scores were often fully or mostly unavailable across all country income levels. Per region, Africa comprised most fully unavailable actions (15·7% [95%-CI 0·0-54·2]) (**Figure 3**), followed by Asia and the Pacific (6·7% [95%-CI 0·0-34·3]). Overall, action availability scores were inversely correlated with country gross national incomes (**Figure 4**).

Figure 3: Action availability per type and region



When recommended actions were mapped back to corresponding guidelines, guideline applicability decreased with World Bank income levels (**Figure 5**). On average, guidelines comprised 1·4% (95%-CI 0·0-33·4) actions fully unavailable in HICs, 3·3% (95%-CI 0·0-73·3) actions fully unavailable in UMICs, 8·0% (95%-CI 0·4-15·6) actions fully unavailable in LMICs, and 22·4% (95%-CI 0·0-50·3) actions fully unavailable in LICs.

Figure 4: Action availability per country gross national income

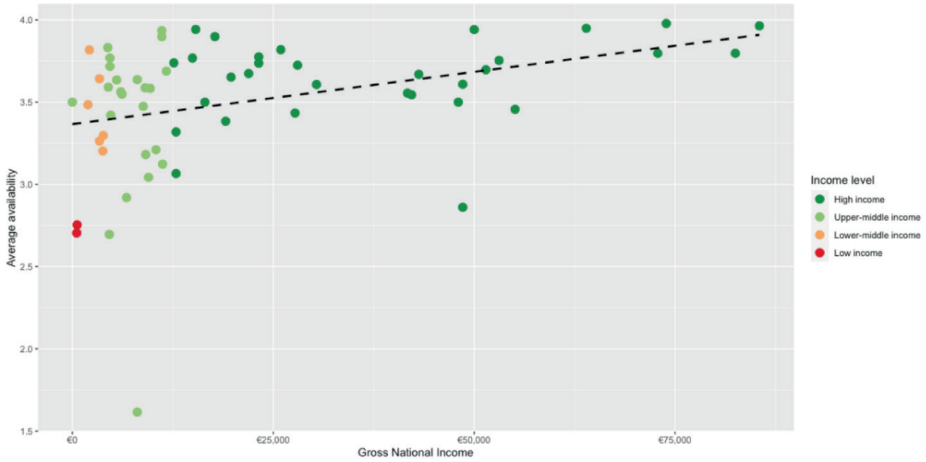
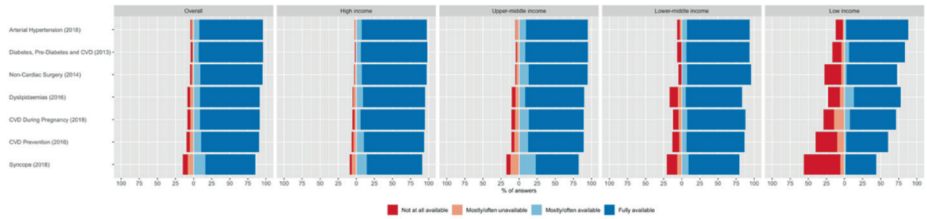


Figure 5: Guideline applicability per country income level

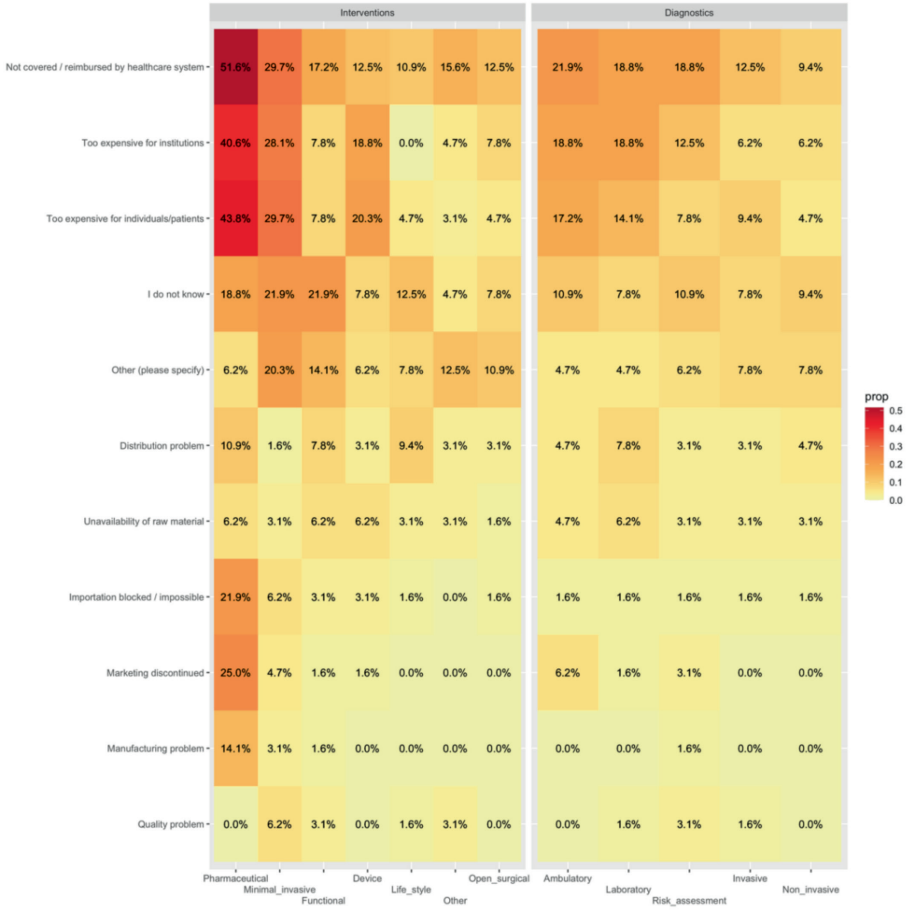


Reasons for action type group inapplicability

A total of 664 reasons for actions being unavailable were reported, matching ten responses per country in total and one per recommendation type (**Figure 6**).

Recommended actions were most often unavailable due to financial reasons, either being not covered or reimbursed by the healthcare system, being too expensive for institutions, or patients. Most responses were received for pharmaceutical interventions (n=150). Most often named reason for pharmaceutical interventions being unavailable was the action not being covered or reimbursed by the healthcare system (n=32), followed by too expensive for patients (n=27), and too expensive for institutions (n=25).

Figure 6: Reasons for action type groups unavailability (percentage of respondents, n=64)



DISCUSSION

In this study we assessed global availability of actions recommended by current ESC guidelines on general cardiology. Moreover, we used this availability to derive global applicability of current ESC guidelines on several general cardiology areas.

We found that recommended actions, availability and applicability of ESC guidelines on general cardiology decreases with country income levels. Notably, some actions proved to be poorly available independent of country income level. Reasons for recommended actions not being available were primarily financial. In particular, for pharmaceutical

interventions financial barriers were often indicated to limit availability. Collectively, these findings show that the clinical practice guidelines of the ESC on general cardiology are not globally applicable and cannot be expected to be implementable by default. Furthermore, some actions recommended by guidelines are unavailable regardless of income level.

The unequal global distribution of medical resources is well documented in literature.[10] In 2013 a systematic review on the distribution of trials globally found that 0·15% of trials were conducted in low-income countries.[11] In like manner, a joint working group of multiple large international public health organizations recently concluded that published research on cardiovascular risks was often inapplicable in lower income countries.[12] Literature on cardiovascular guidelines shows similar patterns.[5] A systematic review in low and lower-middle income countries (LLMICs) from 2016 found just a single guideline published on hypertension in low-income countries and just a few in middle-income countries, most of these not peer-reviewed.[13] Alike, guidelines on stroke, dyslipidemia and risk scores often do not to exist or are deemed inapplicable in LLMICs.[14–17]

The present study shows that despite the aspiration to develop guidelines that are globally applicable, overall, cardiovascular guidelines do not meet this standard. With an estimated 90% of global CVD burden placed in LMICs representing less than 10% of global resources, it is important to support these countries with adequate strategies to reduce and manage CVD risks.[1] Many authors and several large organizations, including the World Health Organization (WHO), regularly call for internationally applicable guidelines. However, although this call is often acknowledged, it remains mainly expressed in words rather than actions.[18] Positive exemptions, putting words into practice, are the World Heart Federation (WHF) and World Gastroenterology Organization (WGO), both actively publishing and applying new methods to improve worldwide applicability of their guidelines.[4,19]

Current cardiovascular guidelines often offer best evidence recommendations only.[20] In LICs, however, applying best evidence is often not possible, as applying best evidence often also bears more costs.[21] Globally applicable guidelines would therefore benefit from providing second and third best recommendations, all becoming first best when higher level bests are not available.[22] In addition, including cost-benefit trade-offs in guideline development processes to assess whether costs are in proportion to health benefits gained might help guidelines to become more globally applicable.

Ethically, the question is whether the guidance developed in HICs is also the best standard for the standard of care in LLIMCs. A globally applicable standard does not mean that everyone should be treated in the same way, but that morally relevant differences are allowed.[28] If the standard of a HIC will never be feasible nor sustainable in an LMIC due to infrastructure or local health care norms, the standard of care of the HIC may not be the best standard for the LMIC. At the same time, in order to determine what is the best standard in a particular setting populations and communities should be engaged. The best standard of care cannot be determined only by members of HICs present in the ESC. A challenge in this regard is the absence of professional medical associations in many LLMICs.[23] For example, the ESC only counts three NCSs from LICs (of whom two responded to the survey for this article) among the 28 LICs worldwide.[9,24] The number of NCS members from LICs remains low, despite extensive recruiting efforts of the ESC Global Affairs Committee in recent years.[25]

Solid cardiovascular guidelines delineate clinical practice in HICs today, leading to better patient outcomes.[2] By contrast, the absence of applicable guidelines in LLMICs hinder these countries to lower their CVD burden at the same rate as higher income countries. Efforts to transfer the large impact that cardiovascular guidelines have in HICs to LLMICs need more attention to lower the global burden of cardiovascular disease.

We recommend that the ESC and other HIC-based developers of guidelines intended to be used globally should consider developing context-stratified recommendations, based on predictable contextual barriers of implementation such as human and financial resources for health.[21] Stratification of guidelines should also consider the level of care in which services are being provided. As indicated by the absence of NCSs in most LICs, many services provided by cardiologists in HICs are provided in LICs by non-specialists or non-physicians, often in primary health care facilities.[26,27]

Study limitations

To appreciate the findings of the current study, some limitations need to be taken into account.

First, only guidelines on general cardiology areas were included, limiting the conclusions of the current work to general cardiology recommendations issued by the ESC. For more specialistic guidelines such as the guideline on coronary revascularization the applicability

might well be (substantially) lower than found in the current study, both overall and across country income levels.

Second, the disseminated survey for this study regarded recommended actions and not recommendations. Therefore, correlations between recommended action availability, recommendation classes, and LoEs could not be assessed.

Third, the disseminated survey was often completed only by one national ESC NCS representative. Unclear was to what extent the individual completing the survey was representative of the NCSs view as a whole on the availability on one or more of the recommended actions. As a result, individual country scores might have been too positive or negative in estimating the applicability of recommended actions.

Fourth, low-income countries were substantially underrepresented in the number of survey responses for this study. The low number of responses was attributable to underrepresentation of LICs in ESC member associations due to absence of professional medical associations in many developing countries.

Fifth and final, NCS representatives often provided less than three reasons for actions being unavailable. Moreover, unavailability reasons were not ranked based on relevance (i.e., most relevant reason listed first). The data on reasons for action unavailability should therefore be interpreted with caution.

CONCLUSIONS

Applicability of cardiovascular guidelines on general cardiology area by the ESC differs across recommended actions, guidelines, countries, and country income levels, with decreasing applicability with decreasing country World Bank GNI levels. Moreover, ESC guidelines comprise multiple recommendations that are limited in their applicability independent of a country's income level. Context-stratifying guideline recommendations to resource levels might help make guidelines more applicable globally.

REFERENCES

1. Murray CJL, Aravkin AY, Zheng P, *et al.* Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 2020; 396: 1223–49.
2. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet* 2017; 390: 415–23.
3. Armstrong PW. Do guidelines influence practice? *Heart* 2003; 89: 349–52.
4. Fried M, Quigley EMM, Hunt RH, *et al.* Are global guidelines desirable, feasible and necessary? *Nat Clin Pract Gastroenterol Hepatol* 2008; 5: 2–3.
5. Wong ND, Morany AE. The U.S. prevention of cardiovascular disease guidelines and implications for implementation in LMIC. *Glob Heart* 2014; 9: 445–55.
6. European Society of Cardiology Member and Affiliated National Cardiac Societies. https://www.escardio.org/The-ESC/Member-National-Cardiac-Societies (accessed Nov 17, 2020).
7. Fanaroff AC, Califf RM, Windecker S, Smith Jr. S, Lopes, Renato D. Levels of Evidence Supporting American College of Cardiology/American Heart Association and European Society of Cardiology Guidelines, 2008-2018. *JAMA* 2019; 27715: 1069–80.
8. van Dijk WB, Grobbee DE, de Vries MC, Groenwold RHH, van der Graaf R, Schuit E. A systematic breakdown of the levels of evidence supporting the European Society of Cardiology guidelines. *Eur J Prev Cardiol* 2019; 26: 1944–52.
9. Serajuddin U, Hamadeh N. New World Bank country classifications by income level: 2020-2021. 2020; published online July 1. https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2020-2021 (accessed July 27, 2021).
10. Puchalski Ritchie LM, Khan S, Moore JE, *et al.* Low- and middle-income countries face many common barriers to implementation of maternal health evidence products. *J Clin Epidemiol* 2016; 76: 229–37.
11. Heneghan C, Blacklock C, Perera R, *et al.* Evidence for non-communicable diseases: Analysis of Cochrane reviews and randomised trials by World Bank classification. *BMJ Open* 2013; 3.
12. Mancia G, Oparil S, Whelton PK, *et al.* The technical report on sodiumintake and cardiovascular disease in low- and middleincome countries by the joint working group of theWorld Heart Federation, the European Society of Hypertension and the European Public Health Association. *Eur Heart J* 2017; 38: 712–9.
13. Owolabi M, Olowoyo P, Miranda JJ, *et al.* Gaps in Hypertension Guidelines in Low- and Middle-Income Versus High-Income Countries: A Systematic Review. *Hypertension* 2016; 68: 1328–37.
14. Bayona H, Owolabi M, Feng W, *et al.* A systematic comparison of key features of ischemic stroke prevention guidelines in low- and middle-income vs. high-income countries. J. Neurol. Sci. 2017; 375: 360–6.
15. Alshamiri M, Ghanaim MMA, Barter P, *et al.* Expert opinion on the applicability of dyslipidemia guidelines in Asia and the middle east. *Int J Gen Med* 2018; 11: 313–22.
16. Kaptoge S, Pennells L, De Bacquer D, *et al.* World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *Lancet Glob Heal* 2019; 7: e1332–45.
17. Khanji MY, Stone IS, James W-Y, *et al.* The applicability of current global cardiovascular risk scores and cardiovascular surrogates in chronic obstructive pulmonary disease. *J Cardiovasc Magn Reson* 2016; 18: 1–2.
18. Olayemi E, Asare E V., Benneh-Akwasi Kuma AA. Guidelines in lower-middle income countries. *Br J Haematol* 2017; 177: 846–54.
19. WHF Global Roadmaps - CVD Roadmaps. https://world-heart-federation.org/cvd-roadmaps/ (accessed July 7, 2021).
20. European Society of Cardiology. Governing Policies and Procedures for the Writing of ESC Clinical Practice Guidelines. 2017.

21. Maaløe N, Ørtved AMR, Sørensen JB, *et al.* The injustice of unfit clinical practice guidelines in low-resource realities. *Lancet Glob Heal* 2021; 9: e875–9.

22. Fried M, Krabshuis J. Can 'Cascades' make guidelines global? *J Eval Clin Pract* 2008; 14: 874–9.

23. Sriram V, Bennett S. Strengthening medical specialisation policy in low-income and middle-income countries. *BMJ Glob Heal* 2020; 5: e002053.

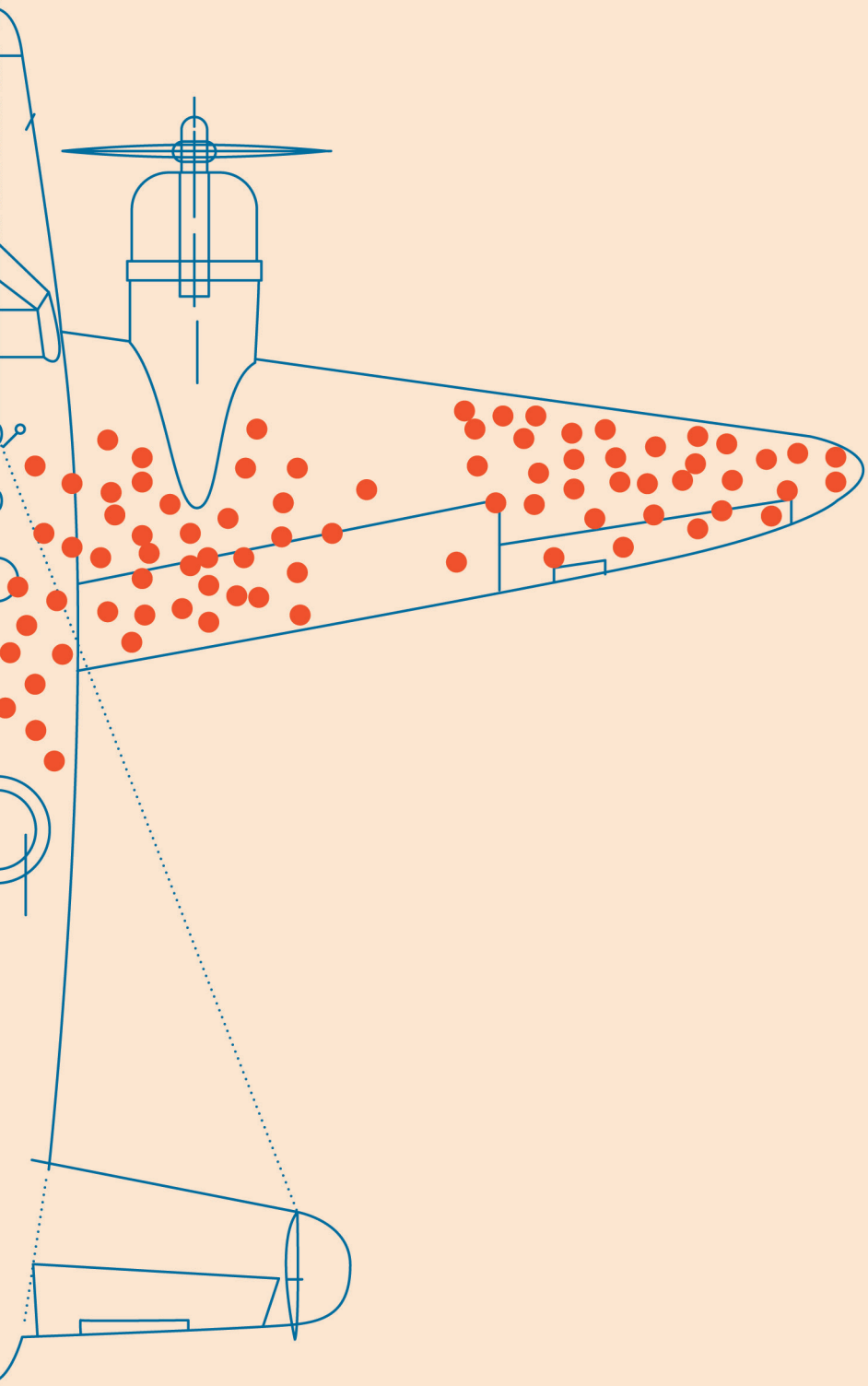
24. European Society of Cardiology. ESC National Cardiac Societies. https://www.escardio.org/The-ESC/Member-National-Cardiac-Societies (accessed July 27, 2021).

25. Grobbee DE. The ESC Global Affairs Committee (2018–20): tackling cardiovascular disease worldwide. *Cardiovasc Res* 2020; 116: e212–4.

26. Tesema AG, Ajisegiri WS, Abimbola S, *et al.* How well are non-communicable disease services being integrated into primary health care in Africa: A review of progress against World Health Organization's African regional targets. *PLoS One* 2020; 15: e0240984.

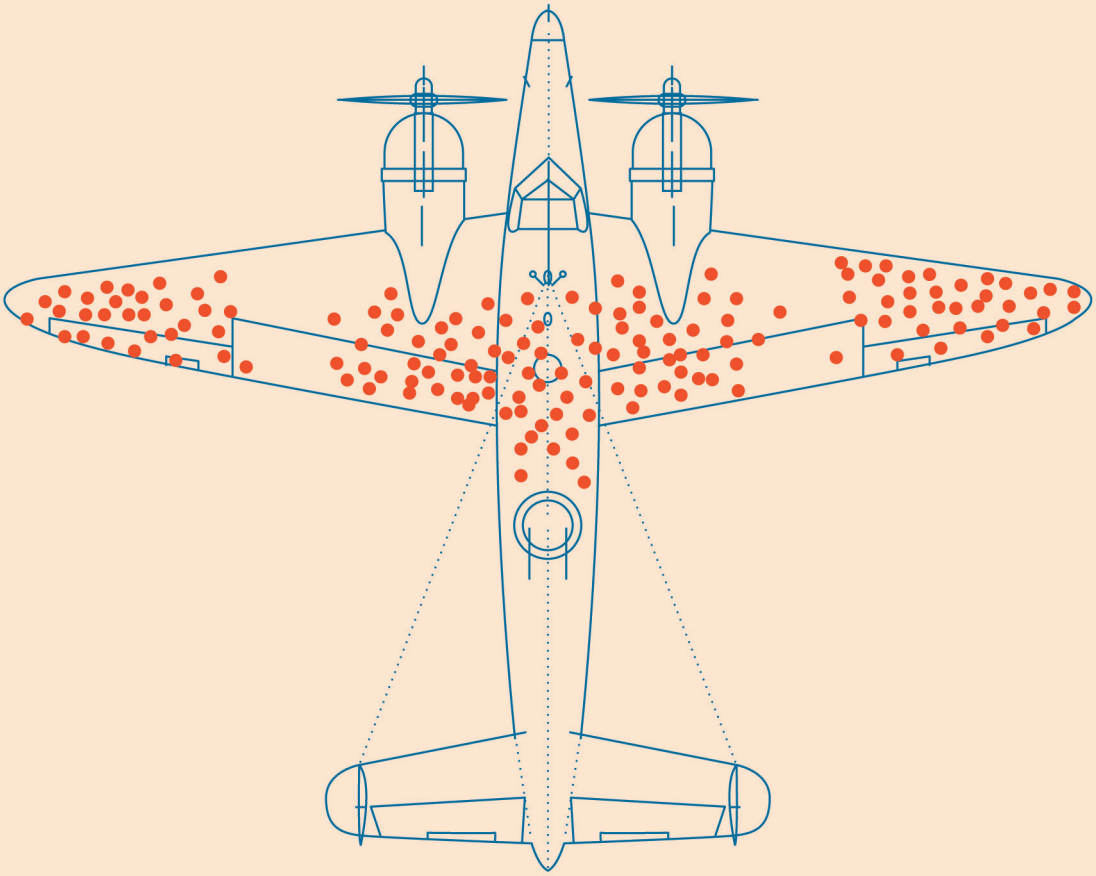
27. Ogedegbe G, Gyamfi J, Plange-Rhule J, *et al.* Task shifting interventions for cardiovascular risk reduction in low-income and middle-income countries: a systematic review of randomised controlled trials. *BMJ Open* 2014; 4: e005983.

28. Graaf R van der, Delden JJM van. What is the Best Standard for the Standard of Care in Clinical Research? *Am. J. Bioeth.* 2009; 9(3):25-43.



PART 2

A bird's eye view on learning healthcare systems



Chapter 4

Learning Healthcare Systems: a scoping review of a methodological and ethical roadmap for implementation

Wouter B van Dijk*, Sara J M Laurijssen*, Rieke van der Graaf, Rolf H H Groenwold, Diederick E Grobbee, Martine C de Vries, Ewoud Schuit

Submitted.

* Authors contributed equally

ABSTRACT

Background

The concept of a Learning Healthcare System (LHS) is promising, but a clear roadmap towards implementation of such a system is missing still. This scoping review aimed to create an overview of methodological and ethical recommendations on the four characteristics the Institute of Medicine defined for an LHS.

Methods

We searched PubMed/MEDLINE, EMBASE and Web of Science up to April 3rd, 2020 for research articles that contained terms related to a Learning Healthcare System.

Results

The search yielded 1,118 unique articles, of which 53 studies were included after full-text assessment. Different recommendations were identified. Methodological recommendations included methods to store data, for quality improvement, and for leadership via stewardship. Ethical recommendations included ideas on informed consent, ethical review and moral obligations for stakeholders.

Conclusion

Our review shows that while methodological and ethical recommendations for LHSs are available, these are fragmented and an integrated approach is missing, thereby limiting the progression towards an LHS.

INTRODUCTION

In 2007 the Institute of Medicine (IoM) suggested that medical practices should become Learning Healthcare Systems (LHSs).[1] Within LHSs clinical care and research are aligned by the use of routinely collected patient data to generate and apply evidence-based medicine and to lower the costs of care while increasing the value of generated evidence. To guide the introduction of LHSs, the IoM defined characteristics and sub-characteristics (**Table 1**) that can function as a roadmap towards an LHS.[1]

Since 2007, different care practices initiated changes to transition into LHSs. In 2021, a review was conducted to investigate the progress of the transition from clinical care towards an LHS. The authors found that while there are different opportunities to learn from healthcare data and improve care, the promise of an LHS has yet to be fulfilled since they find limited evidence of effective system level processes and delivery.[2]

A second review on LHSs described ethical challenges that practices have to overcome when transgressing into LHSs.[3-5] The abrogation of the traditional dichotomy between clinical care and medical research raises questions on topics such as ethical review, informed consent, and privac.[5] A review study on ethical challenges regarding the learning cycle of LHSs identified 67 distinct ethical issues in a total of 65 publications.[5]

Both reviews show that, despite the presence of the roadmap as introduced by IoM, transgression towards LHSs is not easy, which is possibly due to fragmented reporting in favor of certain LHS characteristics as opposed to others, resulting in missing cohesion, contradictions, and ambiguities in methodological and ethical recommendations on LHSs.[5-6] Since there might be ambiguities in methodological and ethical recommendations a clear overview of what is currently recommended by authors is needed. We conducted a scoping review of the literature to create an overview of recommendations on the four characteristics and their sub-characteristics of the IoM for LHSs from both theoretical and applied literature on methods and ethics.

Table 1: Overview of characteristics of LHSs as given by the IoM

Characteristic	Sub-characteristic	Subjects
Science and informatics	Real time access to knowledge	The system continuously and reliably captures, curates and delivers the best available evidence to guide and improve clinical decision-making and healthcare safety and quality
	Digital capture of the care experience	The system captures the care experience on digital platforms for real-time generation and application of knowledge for care improvement
Patient- clinician partnerships	Engaged/ empowered patients	The system focuses on patient needs and perspectives and promotes the inclusion of patients, families, and other caregivers as vital members of the continuously learning care team
Incentives	Incentives aligned for value	The system actively aligns incentives to encourage continuous improvement, identify and reduce waste and reward high-value care
	Full transparency	The system systematically monitors the safety, quality, processes, prices, costs, and outcomes of care, and makes information available for care improvement and informed choices and decision-making by clinicians, patients and their families
Continuous Learning Culture	Leadership instilled culture of learning	The system has leadership committed to a culture of teamwork, collaboration, and adaptability in support of continuous learning as a core aim
	Supportive system competencies	The system constantly refines complex care operations and processes through ongoing team training and skill building; systems analysis and information development; and creation of the feedback loops for continuous learning and system improvement

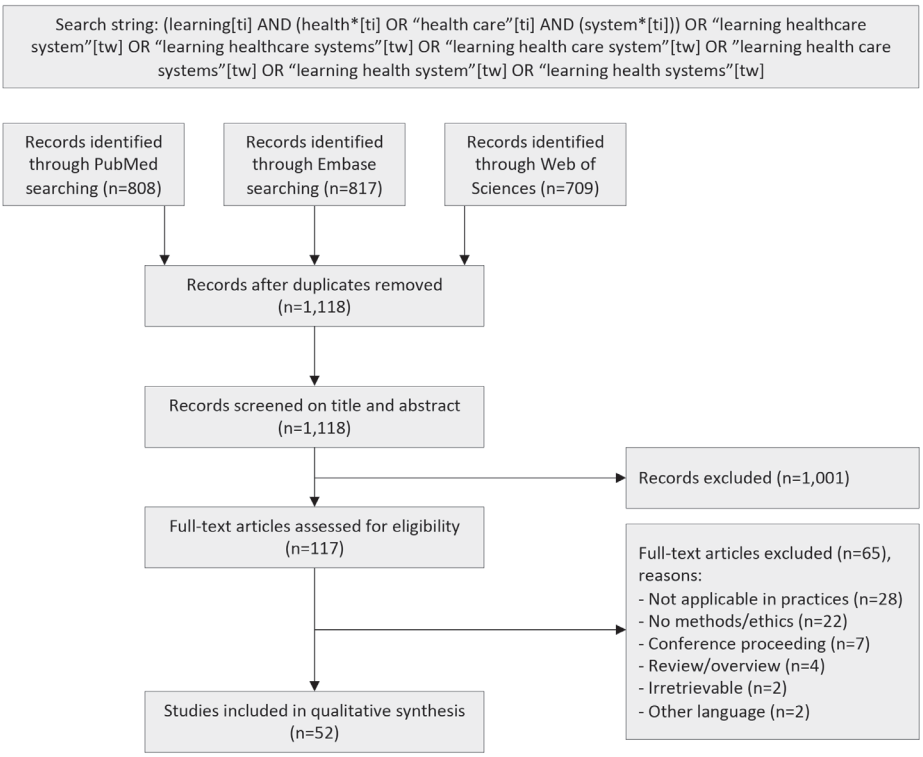
METHODS

This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for scoping reviews ((PRISMA) were applicable.[7] We have included this document in **appendix 1 of the Supplementary Materials**.

Search strategy

A literature search of PubMed, Embase and Web of Sciences was developed and executed on April 3rd, 2020. To identify research articles a search query was built to identify all articles relating to LHSs that contained 'learning health(care) systems' and related terms. The search was set from 2007 since the concept of an LHS was introduced by the IoM in 2007. The search query can be found in **Figure 1**.

Figure 1: Flowchart of applied search strategy



After removal of duplicates, studies were assessed for inclusion based on title and abstract independently by two reviewers (SL and WD). Subsequently, full text articles were assessed by the same two reviewers. Any discrepancies between the reviewers during the screening process regarding the eligibility of articles were resolved by discussion and if needed a third author (ES) was consulted.

Selection criteria

To be eligible, research articles had to (1) describe either methodological and/or ethical recommendations or challenges for LHSs, and (2) the used definition of an LHS had to comply with that of the IoM. Articles not written in English were excluded.

Data extraction and analyses

Articles were labelled based on whether they focused on methodological, ethical or both aspects of an LHS. To be able to identify trends in literature on crucial areas for LHSs, data extraction and analysis were aligned with the LHS characteristics and sub-characteristics

as described by the IoM as backbone for LHSs in their 2012 roundtable (**Table 1**).[1] Articles were grouped on the basis of the characteristics and their sub-characteristics of LHSs by two authors (SL and WD). Subsequently, main themes and topics described in the articles were identified and reported jointly. Roundtable articles of the IoM and systematic reviews were not considered for inclusion but instead were used as reference documents to structure this paper and to validate our search results.[5-6]

To achieve consistent data extraction, a standardized data extraction form was developed, piloted, modified and finalized after discussion amongst the three reviewers (SL, WD, ES).

RESULTS

After removing duplicates, 1,118 articles were screened on title and abstract leaving 117 articles to be screened full text. In total, 53 studies were included after full text assessment (**Figure 1**, **Figure 2**, **Table S1 of the Supplementary Materials**). Hereafter, the results are presented in line with the classification of characteristics and sub-characteristics of an LHS as proposed by the IoM. Only few articles reported recommendations in line with the four categories identified by the IoM. Articles that provided recommendations but did not link to specific LHS characteristics were assigned to LHS characteristics by the authors.

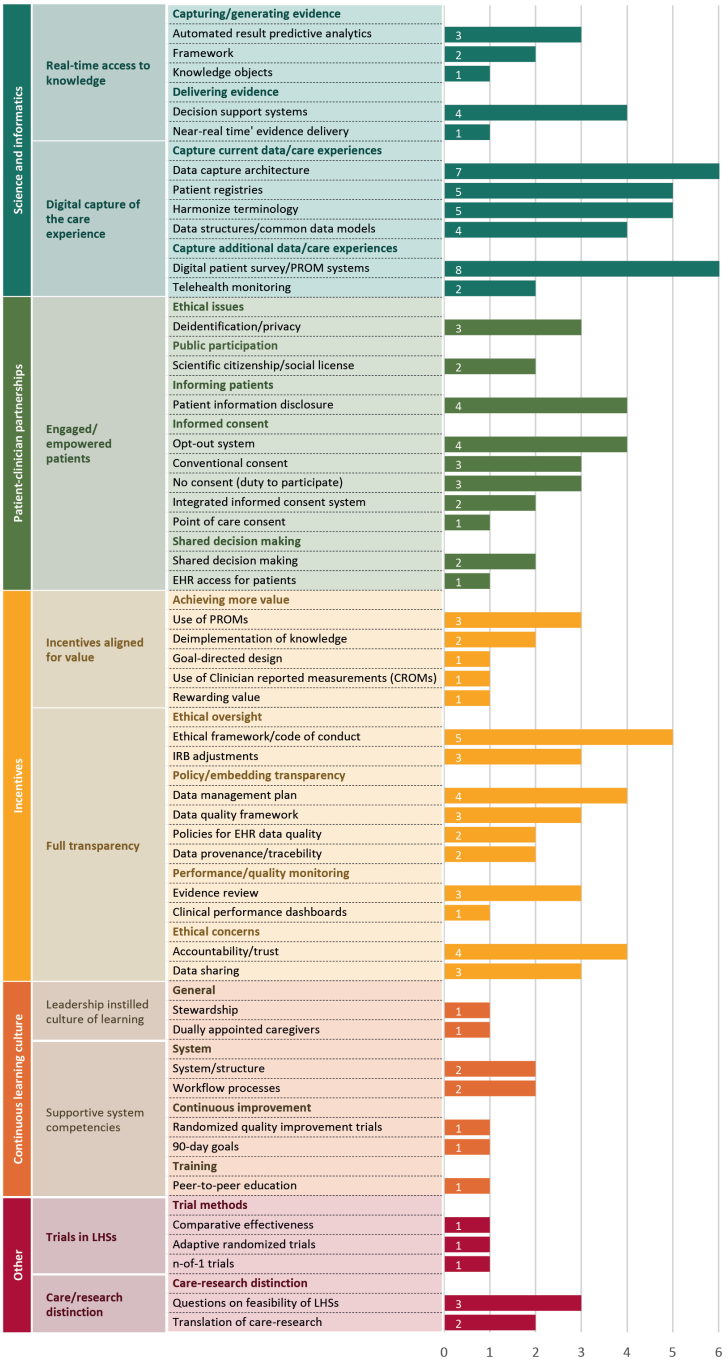
Science and informatics

Real-time access to knowledge

Ten articles reported on methods to leverage science and informatics for real-time access to knowledge.[4,5,8-15] Not being mutually exclusive, six reported on capturing (i.e. generating) [5,8,10,12,13,15] and seven on delivering evidence.[4,5,9,11,13-15]

On evidence capture, one article described a conceptual framework to enable real-time access to data.[8] Two articles mentioned the importance of automating information generation and predictive analytics[10,15]and one described an implemented use cases.[13] Additionally, two articles described the concept of knowledge objects, i.e. frameworks to identify relevant components to gain knowledge from, as implementation of automated analytics.[5,12]

Figure 2: Overview of number of articles per characteristic, sub-characteristic, topic and subject



Regarding evidence delivery, three articles described methods or use cases for decision support systems[5,11,13] and one described methods to incorporate clinical practice guidelines in EHRs.[9] Moreover, one article chronicled a concept combining evidence capture and delivery in an integrated system generating 'near real-time' answers to clinical questions from routinely collected care data in the form of a 'Green button'.[14]

Digital capture of the care experience

Seventeen articles reported on methods to leverage science and informatics for digital capture of the care experience[8,9,16-19], an additional three on ethical issues[15,20,21]. Seven articles described architectures to digitally capture care experiences, commonly by linking existing data sources in data warehouses and integrated repositories.[8,9,11,17-19,2,23] An additional four articles described how to structure data in common data models and make them interoperable.[22,24] Various methods on how to harmonize terminology of large quantities of (routinely) collected data were described in five articles.[22-24,27,28] Five articles mentioned examples of registries already used in practice for LHSs[15,16,17,29,30], one article described linkage of multiple data sources[20] and described the inclusion of health insurance claims data for LHSs[20].

Methods to capture additional care experiences from patients were mentioned in six articles.[9,23,26,27,21,32] Three of these articles described digital patient survey systems to collect these data.[9,31,32] Five described survey systems specifically aimed at collecting patient-reported outcome measures (PROMs) of patients.[9,23,26,31,32] Also, two articles mentioned use of telehealth as a method to collect additional data.[9,15]

Comprising digital capture of the care experience, two articles mentioned the ethical importance of de-identification of data used in LHSs.[8,15] These articles describe that data needs to be deidentified in order to protect the privacy of the patients whose data are being used. Moreover, one article reported concerns of clinicians on de-identification and privacy from clinicians[21].

Patient-clinician partnerships

Engaged, empowered patients

Seventeen articles reported on the engagement and empowerment of patients.[3,4,8,9,15,21,33-40] One article argued that there should be public engagement when EHR data are used in order to ensure social license before usage.[29] Another argued that scientific citizenship is needed before this social license can be established.[4] Both articles recommend that

the general public needs to be educated on the use of their data (on implementations and implications) before their involvement is asked for the LHS.

One of the main ethical topics was the need to inform patients and acquire their informed consent.[3,8,21,29,33-37,39-41] We found different views on how informed consent should be designed for LHSs. These views depend on whether or not articles argued for a moral duty to participate in LHSs for patients. If this moral duty existed, several suggestions were made: not ask consent for certain learning activities within LHSs[21,33,42], design an integrated informed consent system[4,42], Point of Care consent[15] or an opt-out system[8,29,34,41]. When authors did not recognize a duty to participate suggested designs for consent were more conventional[21,35,36]. Empirical articles found that patients wish to be informed on their data usage[35-37], even when these data are de-identified[39.] These empirical studies recommend that patients are at least informed on how their data are being used in the LHS.

Three articles reported on patient engagement in practice: as shared decision making[15,33], to enable patients to email their caregivers[9], or to allow them to download their EHR[9].

Incentives (achieving and rewarding)

Incentives aligned for value

Eight articles reported methods for incentives aligned for value.[9,11,17,22,24,31,32,43] Seven articles suggested methods to achieve higher value in the LHS. These suggestions consisted of: goal-directed design[32], the use of Patient-Reported Outcome Measurements (PROMS) [9,11,17], the use of caregiver reported measurements [31] and de-implementation/ forgetting as a method to eliminate outdated knowledge generated within the LHS[22,43]. One article discussed an incentive to reward value; salaries could be tied to quality improvement outcomes[9].

Full transparency

Full transparency was discussed in 27 articles.[3,4,9,12,15,18,19,21,24,26,29,30,33,36,38,41,44-49] Nine ethics articles reported on ethical oversight methods of care and research activities to ensure patient safety within LHSs[4,29,30,33,36,43,44,46,47]. Different methods for ethical oversight were suggested consisting of; an ethical framework for stakeholders that incorporated moral obligations as part of an ethical LHS[3,33,36,43,36], a federal Institutional Review Board (IRB)[30], a review board specifically designed for LHS[4], and a review board that would safeguard clinical equipoise[47]. Controversy existed regarding the ethical framework suggested by Faden et al., some supported it[21,34,36,40,42], while others were against the

use of such a framework[3,20]. This framework describes moral duties for the different stakeholders involved in an LHS varying from the duty to protect patients to a duty for patients to participate in learning activities[21]. In addition to the framework, a code of conduct for data research was suggested[29].

Ten articles described or suggested policy methods to embed full transparency in LHS s[15,18,19,21,24,26,29,33,38,41] One article described the system they developed to ensure data and privacy protection by design[33] and another described how to protect privacy beyond solely technological methods, i.e. add penalization and data transparency next to data obfuscation[38]. Apparently, oncologists currently have limited trust in these designs still[21], especially when data are shared with third parties such as insurance companies. Processes to assure data quality were discussed in three articles[19,24,26], of which one described comprehensive methods to assess data quality[24]. One article named national standards as a solution to ensure high-quality data[29]. Two articles also mentioned the importance of data provenance and traceability to ensure trust, accountability and transparency in LHSs[18,41]. One article described methods to review evidence generated by LHSs to ensure its quality[12,45,] and one mentioned the risks of labelling research performed in LHSs as 'standard-of-care'-research too easy[45]. This last article recommended that when an LHS is implemented standard of care research should be defined narrowly in order to prevent under regulation.

On a system level, one article described efforts to achieve full transparency among their employees using clinical performance dashboards[9]. Moreover, multiple articles described regular peer-review publishing as ways to enact transparency[9,20].

Different concerns were expressed. Data sharing with stakeholders outside the LHS was considered a concern[21] as well as the trust needed for this data sharing[15,42,49]. An ethical article described that LHSs have a responsibility towards their participants to implement research outcomes within the care that they provide, this responsibility was described as accountability[41].

Continuous learning culture

Leadership-instilled culture of learning

Two articles discussed the leadership instilled culture of learning[9,46]. One article, focusing on ethical aspects of an LHS, described that all stakeholders within the LHS, not only patients, should have the moral responsibility of stewardship[46]. The other, a more methods

oriented article, recognized this stewardship and described dually appointed caregivers (i.e. caregivers with an additional organizational role) and a culture of strong academic affiliation[9]. This article recommended that LHSs should aim to publish findings and install leaders that can oversee academic research[9].

Supportive system competences

We identified seven articles on supportive system competencies[9,17,20,22,25,50,51]. One article gave an integrated framework comprising goals, science, social, ethics and technical dimensions[50]. Another article reported the use of a node structure to stimulate a continuous learning culture in their LHS[17]. In a third system, a more descriptive focus was mentioned in the form of 'emphasis on system redesign' and LEAN methods[9]. Two articles described how workflows or processes could be adjusted to benefit LHSs[25,51]. One of these articles chronicled how this aided the usage of common data models[25] and the other how this could help to automate health data linkage[51].

Two articles described practical methods to incrementally improve care and research in their LHS[17,20]. The first showed how randomized quality improvement experiments improved clinical practice[20] and the second described a method using 90-day goals to make system improvements as a whole[17]. Setting time-specific goals for quality improvement can help to optimize care in an LHS.

Another topic of discussion was the training of those involved in an LHS[9,17]. Training could, for instance, consist of peer-to-peer education[17]. In addition, an article recommended hiring professionals with different expertise that exist outside LHSs to achieve a wide skill set[17]. This can help LHSs to become better integrated within hospital, academia and industry.

Other

Articles not fitting into the characteristics of LHSs as envisioned by the IoM yet describing LHS aspects fell into two categories: new methods for clinical trials and reflections on the care-research distinction.

Trial methods

Three articles reported on trial methodology in LHSs[30,47,52]. One article described various trial designs that can be used for comparative effectiveness studies[52]. Moreover, they visualize integrated comparative effectiveness research wherein randomized and

observational studies are combined. Another article mentioned n-of-1 trials as a trial design of interest[30].

However, some concerns were described regarding the potential lack of clinical equipoise in trials in LHSs[47].

Care-research distinction
Five articles argued that the distinction between healthcare and medical research might not be so evident as the vision of the IoM suggests.[9,53-56] Three of these articles were ethical articles that question the possibility of an LHS as a concept.[53,55,56]

DISCUSSION

When the IoM first conceptualized LHSs, it gave four characteristics and seven sub-characteristics it determined essential for LHSs. Although some authors do refer to these characteristics, our scoping review shows that the vast majority of articles have not been written with these characteristics in mind. There are substantial differences between the number of articles on the various LHS characteristics. The characteristics of incentives, achieving and rewarding and science and informatics were often described in the literature, in particular full transparency received much attention in published literature. In contrast, literature on facets of leadership-instilled culture was largely absent. Overall, high coverage of characteristics could often be explained by interest in one particular topic often addressed, e.g. informed consent (18 articles; characteristic: patient-clinician partnerships), digital architecture and infrastructure (15 articles; characteristic: science and informatics) and ethical oversight (13 articles; characteristic: incentives, achieving and rewarding). In addition, we were not able to identify articles that described how methods can be ethically integrated in an LHS, there were no articles that described scientific validity of knowledge gained within the LHS or the social value of this knowledge. In addition, while observational research is of importance in clinical research, it has not yet been addressed in the context of LHSs.

This scopingreview found three contradictions in the literature on LHSs that were not identified in the previous reviews on LHSs.[5,6] First, most literature on LHSs describes methodological and ethical facets that LHSs are in need of, yet most literature does not describe actual functioning LHSs themselves (yet). It therefore appears that most, if not all, initiatives on LHSs are actually precursors of LHSs still. This is also consistent with

our observation that many authors use the concept of LHSs to promote their viewpoints on new methodological or ethical concepts. However, a comprehensive vision on how to realize LHSs can only be found in a limited selection of articles. We encountered many articles on methods on harmonizing terminologies, for instance, yet did not find these very specific for LHSs. Moreover, articles on ethical frameworks and methods were assigned to the characteristic of full transparency after extensive discussion yet can only fit in this category as they are seen as requirements for full-fledged LHSs.

Second, substantive coverage of important aspects of LHSs is often missing. Articles on actual patient engagement, for example, are missing. Also, patient engagement is only described in classical (passive) ways. Participation of patients in the form of thinking along with or steering research were not found, let alone studies on patient involvement in designing LHSs.

Third, identified recommendations were often incompatible and sometimes contradicted each other. Methods of real-time access to knowledge, for example, are incompatible with most forms of informed consent proposed in ethical literature. Real-time access to knowledge/data is only feasible when patients are asked to provide their consent for data usage upon entering an LHS. When informed consent is asked conventionally through letters for individual studies, however, this would become unworkable. Moreover, point-of-care consent is found to be unacceptable by many ethicists.[48] There is, nonetheless, no ethical consensus on what design of (broad) informed consent should be used within LHSs.[48]

There seems to be a conflict of interest between ethics and methodology. While most ethicists describe that they wish to actively include patients in an LHS by asking informed consent, most methods described in the literature fear that by asking consent there will be selection bias. It would be beneficial if the gap between research methods and ethics is breeched. Future research on building LHSs should focus on guidance for practices to become LHSs that achieve both highly needed research and care outcomes and add to the four characteristics and their sub-characteristics of the LHS to create and maintain the inherent value of an LHS. They should do so with both ethics and methods in mind. In particular, research on leadership and accountability is currently lacking. Leadership is, nonetheless, needed to resolve ethical challenges and to decide on contradicting ideas on moving forwards towards an LHS. Moreover, further guidance on methods, ethics and how to implement LHSs in practice would be beneficial to ease transgression of current systems into LHSs.

Study limitations

To adequately understand the results of this study some limitations have to be acknowledged. First, the search was limited to articles that addressed ethical and/or methodological aspects of an LHS. Therefore, articles might have been missed if the reviewers did not recognize articles properly as either ethical or methodological. Articles outside of ethics and methods in healthcare could have been missed since only medical databases and Web of Science were used. These articles might, for instance, contain ideas on policy, management or implementation. Second, most articles did not describe their findings using the LHS characteristics and their sub-characteristics as given by the IoM. Consequently, we had to assign these findings ourselves, which might have led to some misclassification.

CONCLUSIONS

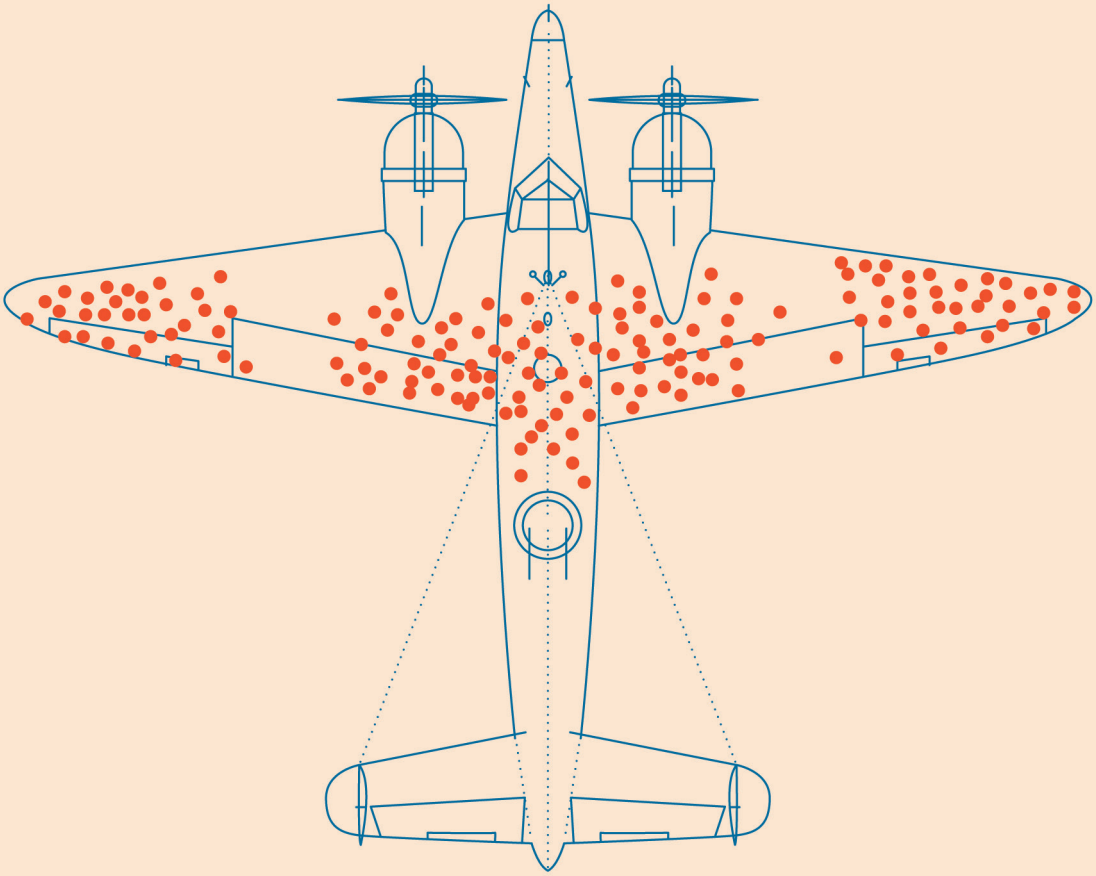
This review shows that even though progress has been made towards LHSs methods and ethics recommendations for LHSs are available for some aspects (e.g. informed consent), but are lacking for other important facets of LHSs (e.g. leadership). On different topics, such as the digital capture of the care experience and patient engagement scholars seem to disagree, leaving practice with little guidance for the actual implementation of LHSs. To proceed towards LHSs, more research is needed to build the constructive roadmap to transform clinical practice into an optimal LHS.

REFERENCES

1. Institute of Medicine. The Learning Healthcare System. Washington, DC: The National Academies Press; 2007.
2. Enticott J, Johnson A, Teede H. Learning health systems using data to drive healthcare improvement and impact: a systematic review. BMC Health Services Research. 2021;21(1):200.
3. Wendler D, Johnson R. When clinical care is like research: the need for review and consent. Theor Med Bioeth. 2016;37(3):193-209.
4. van der Graaf R, Dekking SA, de Vries MC, Zwaan CM, van Delden JJM. Pediatric oncology as a Learning Health System: Ethical implications for best available treatment protocols. Learn Health Syst. 2018;2(2):e10052.
5. Lehmann HP, Downs SM. Desiderata for sharable computable biomedical knowledge for learning health systems. Learn Health Syst. 2018;2(4):e10065.
6. Budrionis A, Bellika JG. The learning healthcare system: where are we now? A systematic review. Journal of biomedical informatics. 2016;64:87-92.
7. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg. 2010;8(5):336-41.
8. Ainsworth J, Buchan I. Combining Health Data Uses to Ignite Health System Learning. Methods Inf Med. 2015;54(6):479-87.
9. Atkins D, Kilbourne AM, Shulkin D. Moving From Discovery to System-Wide Change: The Role of Research in a Learning Health Care System: Experience from Three Decades of Health Systems Research in the Veterans Health Administration. Annu Rev Public Health. 2017;38:467-87.
10. Coiera E. The forgetting health system. Learn Health Syst. 2017;1(4):e10023.
11. Delaney BC, Curcin V, Andreasson A, Arvanitis TN, Bastiaens H, Corrigan D, et al. Translational Medicine and Patient Safety in Europe: TRANSFoRm--Architecture for the Learning Health System in Europe. Biomed Res Int. 2015;2015:961526.
12. Guise JM, Savitz LA, Friedman CP. Mind the Gap: Putting Evidence into Practice in the Era of Learning Health Systems. J Gen Intern Med. 2018;33(12):2237-9.
13. Keim-Malpass J, Kitzmiller RR, Skeeles-Worley A, Lindberg C, Clark MT, Tai R, et al. Advancing Continuous Predictive Analytics Monitoring: Moving from Implementation to Clinical Action in a Learning Health System. Crit Care Nurs Clin North Am. 2018;30(2):273-87.
14. Longhurst CA, Harrington RA, Shah NH. A 'green button' for using aggregate patient data at the point of care. Health Aff (Millwood). 2014;33(7):1229-35.
15. Maddox TM, Albert NM, Borden WB, Curtis LH, Ferguson TB, Jr., Kao DP, et al. The Learning Healthcare System and Cardiovascular Care: A Scientific Statement From the American Heart Association. Circulation. 2017;135(14):e826-e57.
16. Bhandari RP, Feinstein AB, Huestis SE, Krane EJ, Dunn AL, Cohen LL, et al. Pediatric-Collaborative Health Outcomes Information Registry (Peds-CHOIR): A learning health system to guide pediatric pain research and treatment. The Journal of the International Association for the Study of Pain. 2016;157(9):2033-44.
17. Britto MT, Fuller SC, Kaplan HC, Kotagal U, Lannon C, Margolis PA, et al. Using a network organisational architecture to support the development of Learning Healthcare Systems. BMJ Qual Saf. 2018;27(11):937-46.
18. Curcin V. Embedding data provenance into the Learning Health System to facilitate reproducible research. Learn Health Syst. 2017;1(2):e10019.
19. Priest EL, Klekar C, Cantu G, Berryman C, Garinger G, Hall L, et al. Developing electronic data methods infrastructure to participate in collaborative research networks. EGEMS (Wash DC). 2014;2(1):1126.
20. Horwitz LI, Kuznetsova M, Jones SA. Creating a Learning Health System through Rapid-Cycle, Randomized Testing. N Engl J Med. 2019;381(12):1175-9.

21. Mayo RM, Summey JF, Williams JE, Spence RA, Kim S, Jagsi R. Qualitative Study of Oncologists' Views on the CancerLinQ Rapid Learning System. J Oncol Pract. 2017;13(3):e176-e84.
22. Ethier JF, McGilchrist M, Barton A, Cloutier AM, Curcin V, Delaney BC, et al. The TRANSFoRm project: Experience and lessons learned regarding functional and interoperability requirements to support primary care. Learn Health Syst. 2018;2(2):e10037.
23. Harle CA, Lipori G, Hurley RW. Collecting, Integrating, and Disseminating Patient-Reported Outcomes for Research in a Learning Healthcare System. EGEMS (Wash DC). 2016;4(1):1240.
24. Kapsner LA, Kampf MO, Seuchter SA, Kamdje-Wabo G, Gradinger T, Ganslandt T, et al. Moving Towards an EHR Data Quality Framework: The MIRACUM Approach. Stud Health Technol Inform. 2019;267:247-53.
25. Kuchinke W, Karakoyun T, Ohmann C, Arvanitis TN, Taweel A, Delaney BC, et al. Extension of the primary care research object model (PCROM) as clinical research information model (CRIM) for the "learning healthcare system". BMC Med Inform Decis Mak. 2014;14:118.
26. McGlynn EA, Lieu TA, Durham ML, Bauck A, Laws R, Go AS, et al. Developing a data infrastructure for a learning health system: the PORTAL network. J Am Med Inform Assoc. 2014;21(4):596-601.
27. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, et al. PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc. 2014;21(4):602-6.
28. Klann JG, Phillips LC, Turchin A, Weiler S, Mandl KD, Murphy SN. A numerical similarity approach for using retired Current Procedural Terminology (CPT) codes for electronic phenotyping in the Scalable Collaborative Infrastructure for a Learning Health System (SCILHS). BMC Med Inform Decis Mak. 2015;15:104.
29. Eichler HG, Bloechl-Daum B, Broich K, Kyrle PA, Oderkirk J, Rasi G, et al. Data Rich, Information Poor: Can We Use Electronic Health Records to Create a Learning Healthcare System for Pharmaceuticals? Clin Pharmacol Ther. 2019;105(4):912-22.
30. Margolis PA, Peterson LE, Seid M. Collaborative Chronic Care Networks (C3Ns) to transform chronic illness care. Pediatrics. 2013;131 Suppl 4:S219-23.
31. Flum DR, Alfonso-Cristancho R, Devine EB, Devlin A, Farrokhi E, Tarczy-Hornoch P, et al. Implementation of a "real-world" learning health care system: Washington State's Comparative Effectiveness Research Translation Network (CERTAIN). Surgery. 2014;155(5):860-6.
32. Fore D, Goldenhar LM, Margolis PA, Seid M. Using goal-directed design to create a novel system for improving chronic illness care. JMIR Res Protoc. 2013;2(2):e43.
33. Boeckhout M, Scheltens P, Manders P, Smit C, Bredenoord AL, Zielhuis GA. Patients to learn from: on the need for systematic integration of research and care in academic health care. J Clin Transl Res. 2018;3(Suppl 3):401-6.
34. Botkin JR. Transparency and choice in learning healthcare systems. Learn Health Syst. 2018;2(1):e10049.
35. Dranseika V, Piasecki J. Transparent Defaults and Consent for Participation in a Learning Health Care System: An Empirical Study. J Empir Res Hum Res Ethics. 2020;15(4):261-70.
36. Faden RR, Kass NE, Goodman SN, Pronovost P, Tunis S, Beauchamp TL. An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. Hastings Cent Rep. 2013;Spec No:S16-27.
37. Jagsi R, Griffith KA, Sabolch A, Jones R, Spence R, De Vries R, et al. Perspectives of Patients With Cancer on the Ethics of Rapid-Learning Health Systems. J Clin Oncol. 2017;35(20):2315-23.
38. Kasperbauer TJ. Protecting health privacy even when privacy is lost. J Med Ethics. 2020;46(11):768-72.
39. Sacristan JA, Dilla T. No big data without small data: learning health care systems begin and end with the individual patient. Journal of Evaluation in Clinical Practice. 2015;21(6):1014-7.
40. Weinfurt KP, Lin L, Sugarman J. Public views regarding the responsibility of patients, clinicians, and institutions to participate in research in the United States. Clin Trials. 2019;16(6):574-9.

41. Kass NE, Faden RR. Ethics and Learning Health Care: The Essential roles of engagement, transparency, and accountability. Learn Health Syst. 2018;2(4):e10066.
42. Miller FG, Kim SY. Personal Care in Learning Health Care Systems. Kennedy Inst Ethics J. 2015;25(4):419-35.
43. Kass NE, Faden RR, Goodman SN, Pronovost P, Tunis S, Beauchamp TL. The Research-Treatment Distinction: A Problematic Approach for Determining Which Activities Should Have Ethical Oversight. Hastings Center Report. 2013;43:S4-S15.
44. Fiscella K, Tobin JN, Carroll JK, He H, Ogedegbe G. Ethical oversight in quality improvement and quality improvement research: new approaches to promote a learning health care system. BMC Med Ethics. 2015;16(1):63.
45. Kim SYH. Ethical issues in pragmatic trials of "standard-of-care" interventions in learning health care systems. Learn Health Syst. 2018;2(1):e10045.
46. Largent EA, Miller FG, Joffe S. A prescription for ethical learning. Commentary. Hastings Cent Rep. 2013;Spec No:S28-9.
47. London AJ. Learning health systems, clinical equipoise and the ethics of response adaptive randomisation. J Med Ethics. 2018;44(6):409-15.
48. McLennan S, Kahrass H, Wieschowski S, Strech D, Langhof H. The spectrum of ethical issues in a Learning Health Care System: a systematic qualitative review. Int J Qual Health Care. 2018;30(3):161-8.
49. McNolty LA, Payne R. Relying on Trust for Research on Medical Practice in Learning Health Systems. Am J Bioeth. 2015;15(9):30-2.
50. Lessard L, Michalowski W, Fung-Kee-Fung M, Jones L, Grudniewicz A. Architectural frameworks: defining the structures for implementing learning health systems. Implement Sci. 2017;12(1):78.
51. Van Eaton EG, Devlin AB, Devine EB, Flum DR, Tarczy-Hornoch P. Achieving and sustaining automated health data linkages for learning systems: barriers and solutions. EGEMS (Wash DC). 2014;2(2):1069.
52. Webster M, Stewart R, Aagaard N, McArthur C. The learning health system: trial design and participant consent in comparative effectiveness research. Eur Heart J. 2019;40(15):1236-40.
53. Lee SS, Kelley M, Cho MK, Kraft SA, James C, Constantine M, et al. Adrift in the Gray Zone: IRB Perspectives on Research in the Learning Health System. AJOB Empir Bioeth. 2016;7(2):125-34.
54. Piasecki J, Dranseika V. Learning to Regulate Learning Healthcare Systems. Camb Q Healthc Ethics. 2019;28(2):369-77.
55. Rosenthal GE. The role of pragmatic clinical trials in the evolution of learning health systems. Trans Am Clin Climatol Assoc. 2014;125:204-16; discussion 17-8.
56. Selby JV, Krumholz HM. Ethical Oversight: Serving the Best Interests of Patients. Hastings Center Report. 2013;43:S34-S6.



Chapter 5

The social value of knowledge in learning healthcare systems

Wouter B van Dijk, Rieke van der Graaf, Ewoud Schuit, Sara J M Laurijssen, Rolf H H Groenwold, Diederi, Martine C de Vries

Submitted.

ABSTRACT

Learning Healthcare Systems (LHSs) have been proposed to enable the integration of research with health care systems in a way beneficial for both. The idea of LHSs is that by using real patient data researchers can produce studies that are more relevant and more likely to be used by practitioners, creating a strengthened feedback loop. LHSs are therefore expected to increase the social value of the knowledge produced in them. We analyze the social value of research done in LHSs and compare it with the conventional research in medicine.

We distinguish three types of social value of knowledge: (1) internal, (2) external and (3) temporal. Research has internal social value if it is locally applicable. In LHSs the internal social value of research increases as it can be tailored to the wants and needs of LHS populations. Research has external social value if it is applicable in general outside the environment it was generated in, which is at risk for knowledge generated in LHSs. For if research is tailored to LHSs by setting strict eligibility criteria for participants these criteria can limit its applicability outside it. Moreover, recruiting participants solely in LHSs limits the size of study populations, which can limit its statistical accuracy and inhibit the precision of the results. Last but not least, research has temporal social value if it is reliable, being at risk in a LHS as knowledge is implemented faster. When less time exists between generating and implementing research results, less time exists to check the validity these results increasing the risk of changing clinical practice on spurious research.

As knowledge generated in LHSs will have more value locally and possibly be limited externally, research may diverge into different directions. This is problematic when having to choose between results from LHSs or international guidelines. When diverging research is of comparable quality, choices have to be made on which results should prevail. Clinicians, researchers and policy makers working in an LHS should therefore take specific measures to safeguard the societal value of their research, among others advocating awareness, installing independent validation committees and fostering LHS transcending collaborations.

INTRODUCTION

The Institute of Medicine (IoM) introduced the concept of Learning Healthcare Systems (LHSs) to enable the transfer of knowledge between care and research in a timely and affordable manner.[1,2] LHSs could facilitate the translation of ever-increasing amounts of data into knowledge and be used to direct resources spent on research and implementation.[1,2] Four characteristics have been defined as essential for LHSs: (1) incentives aligned on value, (2) a culture of continuous learning, (3) effective use of science and informatics, and (4) effective patient-clinician partnerships.[2]

Current LHS initiatives often try to directly implement these characteristics in real-time data systems to allow faster and more responsive evidence generation.[3] These LHSs could also be described as specific data sources that enable faster dynamic population research studies.[4] Often, these initiatives have a local or regional focus, meaning the evidence generated in the LHS is applicable locally, but may be hard to replicate outside that context.[5,6] LHSs generally focus on two of the characteristics given by the IoM.[3] First, using computing power and routinely collected data more effectively by applying science and informatics, statistical knowledge is built faster. Second, patient-clinician partnerships are stimulated to create research that is more usable in the local context. This second goal is generally materialized by using Patient Reported Outcome Measures (PROMs) to drive local research agendas.[3,7]

In literature the social value of science is often found to be restricted by how fast research is done and by how quickly it is employed in local setting.[8] The premise of LHSs to apply research more quickly has therefore set the expectation that the social value of scientific knowledge will also increase.[2] In this paper, we analyze the social value induced by research in LHSs and their advantages and disadvantages compared to conventional medical research. For this purpose, we distinguish three kinds of social value: (1) internal, (2) external and (3) temporal.[9]

We begin by specifying the elements of research social value found in literature. Then we analyse the effect of LHSs on social value. Afterwards we assess the consequences these effects might change the hierarchy of research. Finally, we reflect on how clinicians, researchers and policy makers should anticipate upon the changed social value of research of data from LHSs.

To illustrate our analysis, we will use the Utrecht Cardiovascular Cohort (UCC) as an example of an LHS (see **Box 1**).

Box 1: The Utrecht Cardiovascular Cohort: an LHS in the Making

The UCC was designed as a prospective cohort study which included patients referred to the cardiology department of the University Medical Center Utrecht with a potential cardiovascular diagnosis.[10] From its founding the UCC collected short-term and long-term data of its participants by collecting data during routine care. From 2017 onwards, the UCC is presenting itself as a learning cardiovascular healthcare system, aiming to assist clinicians with direct feedback to increase guideline compliance and increase patient outcomes. To further achieve this aim, follow-up information is collected from a number of additional sources on all patients in the UCC to eventually be able to establish a continuous learning healthcare system with feedback to clinician's decision support dashboards.[10]

BREAKING DOWN SOCIAL VALUE: THE INTERNAL, EXTERNAL AND TEMPORAL SOCIAL VALUE OF EVIDENCE

In research the concept of `social value` means the prospect of generating knowledge and means necessary to protect and promote people's health.[11] In context of LHSs, social value can be classified into three types: internal, external and temporal (Table 1).[9]

Internal social value is the internal applicability of the generated knowledge. In ethical literature internal social value is often referred to as *responsiveness*.[12] Internal social value represents the extent to which the knowledge is applicable to the population in which the evidence was generated, in biomedical, organizational and financial terms.[13] In addition, availability of interventions and consideration for health priorities of the population are also considered important factors of responsiveness.[12,14] Similar to when multiple studies are conducted in different contexts, results from different LHSs may not be applicable in other LHSs. For example, prediction models may provide valuable insights in characteristics of an LHS population yet will often be of little value outside the LHS.

External social value is the opposite of internal social value and implies the extent to which knowledge is applicable outside an LHS, defined as the applicability of findings in populations other than the population where they were obtained from.[13] In ethical contexts, external applicability is often referred to as *generalizability* or *societal value* and is commonly used as argument to justify allocation of resources.[9,13] As resources are finite, evidence brought forth by research should have societal social value to warrant the use of public resources.[13] As such, generalizability is used as touchstone for societal social value, simply meaning that more should be able to benefit from research results than those who participated in the study only.

Temporal social value is found in the reliability of knowledge. Temporal social value represents the extent to which research findings can stand the test of time.[15,16] In the paradigm of evidence based medicine evidence temporal social value is tested by peer review and replication studies trying to falsify results.[16–18] In this process evidence substantiating knowledge is tested on its study design, execution and found results, to validate the quality of the research and filter out statistical anomalies.[19] Over time the value of evidence can either increase as results are confirmed or decrease when results are refuted by other researchers.

Table 1: Types of social value and their associated elements

Type	Description	Synonyms
Internal social value	The value generated knowledge has for the LHS it was generated in.	Responsiveness; local applicability
External social value	The value generated knowledge has outside the LHS it was generated in, i.e. for society.	Generalizability, external applicability
Temporal social value	The value and reliability generated knowledge proves to have over time.	Reliability, validity

The internal social value of knowledge in LHSs

LHSs are ideally situated to make research results more relevant of for their own populations and practices for three reasons.

First, LHSs are able to generate knowledge faster. Current research findings are estimated to take 17 years before being implemented in practice, making timeliness and relevance of the knowledge at implementation questionable.[20] By using routinely collected data and automated computer analysis, LHSs can generate new insights from information faster. Fast evidence generation increases the timeliness of knowledge and improve its social value.

The Utrecht Cardiovascular Cohort (UCC) adds internal social value of research by having a near real-time development and implementation of decision support systems in clinical dashboards based on EHR data to assist clinicians and their patients in making choices.[21,22]

Second, LHSs are able to generate more locally useable results resulting from easier access to real-world data relevant for their context. Knowledge from current (non-LHS) research, e.g. randomized clinical trials, is frequently found to be limited in its applicability due to artificial participant selection and data collection.[23,24] The use of real-world data increases the applicability of knowledge by matching research more closely to real-world challenges. Pragmatic trials offer insights in the value using real-world data can have for clinical practice.[25] As research performed in LHSs utilizes EHR data and focuses specifically on the population within the LHS itself its researchers tend to encounter less barriers to this data than other systems.[26] Improved access to real-world data can be seen in UCC's dashboards, as they make individual patient data directly visible to clinicians and patients.[22] Eventually, the UCC's goal is to directly draw new scientific insights from EHR data, creating a full-fledged LHS.

Third, LHSs are able to generate more applicable results because research questions can be tailored to its population. When patients are aware that research done in their LHS can also benefit themselves they are more willing to participate in research.[27] Direct access to results from research on PROMS considered relevant on a local level through decision support systems can give clinicians and their patients the ability to directly review prognoses and therapeutic options expressed in measures that matter to patients. When patients find other items more relevant than those shown on the dashboards this can eventually result in changes in the research agenda in LHSs.[1] For patient engagement, the UCC actively seeks advice and expertise on what patients want to gain from research by actively involving them in all projects. Moreover, in the UCC a monthly Cardiovascular Research Café is organized to communicate outcomes, collect feedback from a wider array of patients, and reflect if outcomes (still) align with patients' needs.

The external social value of knowledge in LHSs

The external social value of knowledge is determined by its relevance or external applicability for populations other than in the one where the knowledge was generated from. Knowledge from LHSs is at risk of being limited in its external social value by two factors.

First limiting factor is limitations due to the chosen research domain. Scientific research is expected to be generalizable to the broader population, i.e. domain, from which the study population was randomly sampled.[28] The theoretical research domain is the population researchers want their research to be applicable to.[28] The material research domain is the domain that the study population in practice represent after being sampled. Generalizability of LHS knowledge becomes poor when a theoretical domain is chosen so narrowly that it can only be found in a single LHS or when the theoretical and material domain do not align, when, for example, effect modifiers are (un)consciously included in eligibility criteria. As is common in LHSs, knowledge generation is focused on creating responsive knowledge with high internal social value. The focus on internal social value warrants to be alert on research's generalizability too. Also, strict eligibility criteria make the generated evidence less compatible with evidence generated elsewhere.[29] An effect that is amplified when core outcomes of research are chosen differently between LHSs as result of increased patient involvement.[30]

The second limiting factor is lack of precision of evidence. In research, sample sizes need to be sufficiently large to allow detection of meaningful associations with an adequate amount of precision, i.e. small statistical confidence instervals.[28] Unless effects are large, substantial sample sizes are needed to allow rejection of a null hypothesis of no association, which may practically limit local research. It is currently not uncommon that randomized controlled trials need more than 10.000 patients to have sufficient power to detect treatment benefits with sufficient confidence.[31] Dependent on whether LHSs encompass one hospital department, a whole hospital, a region a country or more, evidence obtained from LHS populations may be too small to provide sufficient research power. Consequently, it may be easier to recruit study participants while it is harder to recruit enough participants for sufficient precision.

Although not unique for LHSs, this need for precision also interferes with the desire for more personalized medicine, since these methods are data hungry and depend on large datasets to facilitate the needed subgroup analyses.[32] Evidence of multiple LHSs might have to be combined to facilitate research with sufficient power and precision. But combining evidence from LHSs requires that findings are generalizable across LHSs. Alignment on domain definition and core outcomes across LHSs is crucial to combine results at a later stage.[33,34]

The temporal social value and reliability of evidence

The way in which knowledge is obtained in LHSs will not only affect the internal and external social value of the knowledge, but also affect its temporal social value or reliability. LHSs provide the opportunity to, faster than currently, apply knowledge in clinical practice. However, faster knowledge generations impact reliability of knowledge as reliability of knowledge builds over time.

In his work on the scientific knowledge filter, Bauer asserted that knowledge generation starts with subjective hypotheses from researchers.[16] These hypotheses become (unpublished) frontier science when worked out. Frontier evidence heavily depends on a researcher's skills and integrity still and history shows that over time many scientific findings are wrong or have to be modified.[17] When deemed worthy by peers, evidence is published as primary literature. Primary literature mainly orders evidence and makes it accessible.[35] Research becomes secondary literature if other researchers find the published evidence useful, valid and cite it or take it up in literature reviews and meta-analyses. Ultimately, evidence becomes textbook or knowledge as it becomes generally accepted. With every step in the knowledge filter chances of knowledge comprising false results decline. The two main factors dominating this filtering process are the passing of time and appraisals from other researchers in the form of replication studies, citations and usage.

Although larger findings in LHSs will always be published still, LHSs also generate evidence less likely to be published, as it is intended for local application only. The UCC for example published about its dashboards and the impact they have on their population yet will not publish about every update of underlying algorithms.[22] When the number of findings increase, and research tasks are automated publishing all findings can become unachievable with a speed of knowledge generation too high to be processed by peer-review journals before being applied in practice. This raises questions whether evidence will become primary literature, and subsequently secondary literature, at all. Moreover, it also raises the question how evidence will and actually can be appraised critically within LHSs.

Evidence from LHSs is at risk of being appraised less by independent researchers when control mechanisms are not embedded in their organizations. Time allows evidence to be thoroughly appraised by peers, in which not just the number of researchers reviewing evidence is of importance, but also the passing of time to allow confirmation or falsification with replication studies. In LHSs the time before implementing evidence can become

short enough that such rigor cannot be applied, and clinical practice can be exposed to increased risks. If the UCC would evaluate a new medical device on EHR data, when could they be certain enough of the validity of the data to implement their findings? Questions like become more prominent as more LHSs take off. In 2019 an article was published showing promising results of multiple randomized quality improvement experiments.[36] Although, the authors would like to publish their results they highlight that this was hard for most of their results.

Conversely, publication of promising yet methodologically limited results can also hamper future higher quality research. To quote Sir Austin Bradford Hill 'not infrequently, clinical workers publish favorable results on three or four cases and conclude their article by suggesting that this is the method of choice, and that what is required now is a trial on an adequate scale. They do not seem to realize that by their very publication they have vastly increased the difficulties of that trial or, indeed, made it impossible.[37] Swedish researchers has shown that implementing new interventions before obtaining sufficiently reliable evidence is available is a realistic threat.[38] They showed how the use of thrombus aspiration for myocardial infarctions rose and fell with trends in evidence as first observational results looked promising and were then falsified by later trials.

The main question for evidence generated in LHSs is: when can the evidence generated within the LHS be applied in clinical practice? Even when only applied locally. Tools like the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework can provide guidance herein. These tools, however, are usually developed for reviewing evidence being primary literature or better and tend to focus on systematic reviews of healthcare interventions.[39] To ensure knowledge generated in LHSs is reliable enough additional safeguards should be built in LHSs. Although these safeguards will not be able to fully compensate for the validation of evidence over time, they will allow validating evidence found in LHSs against evidence from the outside. Safeguards could include incorporating or testing found evidence against existing evidence using, for example, Bayesian statistics or sensitivity analysis, or the use of other evidence synthesis methods before the generated evidence is implemented in clinical practice.[30] Other options could entail periodical publishing of evidence findings, following the example of living reviews, to give the scientific community opportunities to test new findings.[40] Regardless, procedures should be built into LHSs to warrant a systematic review process of evidence found within LHSs. Currently, research is mainly validated and monitored before and sometimes during its conduct by research ethics committees and data and

safety monitoring boards.[41] In LHSs there is a need to install more formal post-research checks to ensure that only the most trusted evidence gets implemented in clinical practice.

CONSEQUENCES OF LHSS FOR THE RESEARCH HIERARCHY

Research results from LHSs can thus be expected to be generated faster, be more responsive and locally applicable. At the same time, LHSs may encompass small or highly specific study populations and settings, which may or may not necessarily affect their generalizability and precision of the knowledge. Consequently, results may be more of direct use to the LHS, rather than to society or even leading to risks when implemented locally.

When we look at how evidence generated in LHSs fits into the current paradigm of evidence-based medicine a paradox is found. Presently, clinical practice is mainly driven by recommendations of national and international clinical practice guidelines.[42] These guidelines classify their evidence into levels of reliability adhering to the classic evidence pyramid. From best to worse evidence is categorized as follows: meta-analyses and large Randomized Controlled Trials (RCTs) (level A), small RCTs and large observational studies (level B) and smaller observational studies, case series and expert opinion (level C). Multiple independent reviews on clinical practice guidelines of various medical fields found that current evidence levels are low in general.[42,43] About 15% of the recommendations in guidelines are grounded on level A evidence, 30% on level B and 55% on level C. Although, these aggregates should be broken down to accurately identify evidence gaps as not all areas required level A evidence always (in therapeutic areas trials are the gold standard but not necessarily in diagnostic research), a large portion remains based on expert opions.[42]

Evidence from LHSs will generally not take the form of level A evidence, will sometimes take the form of level B evidence but most of the times should be taken as level C evidence. This affects the research hierarchy between evidence generated inside and outside LHSs.

Guideline recommendations substantiated by level A evidence will likely prevail over results found in LHSs whether aligning with its results or not. Thus, when level A evidence is present LHSs will limitedly or not change clinical practice. When guideline recommendations are supported by level B evidence, however, it might not be as clear which evidence should prevail. If guideline recommendations are based on the results of an observational study with 3.000 participants, and data from, for example, the UCC shows

opposite results, which results should be followed? Indeed, contrasting evidence is found more often. Currently, however, best practice is to follow guideline recommendations. But does this best practice remain when these recommendations are challenged by results from an LHS? In particular considering that these results are probably more responsive to the specific population.

When guideline recommendations are based on level C evidence (expert opinion) and evidence from an individual LHS suggests otherwise it makes sense to follow the results from the LHS. This would allow LHSs to provide recommendations more tailored to local populations.

DISCUSSION AND MESSAGES FOR CLINICIANS, RESEARCHERS AND POLICY MAKERS

On the scale of single LHSs, knowledge that is generated may be preferable over international guideline recommendations as it enables researchers to answer questions relevant to their local clinical practice faster. In this way LHSs contribute to the imperatives to solve (the informational and the value imperative) set by the IoM.[1,2] On a societal scale, however, LHSs might also bring some challenges, making it harder to determine whether the knowledge generated by LHSs should always be embraced. LHSs will most likely assist in raising evidence levels for clinical practice, yet they require caution with a view to validity and generalizability.

In interpreting and applying results obtained in LHSs it is important to be aware of how knowledge generated by LHSs will be affected and react to changes when they emerge. Apart from being aware of characteristics of the knowledge from LHSs, clinicians, researchers and policy makers working in an LHS can take specific measures to improve the societal value of their research.

First, by creating protocols and procedures to render evidence generated in LHSs as generalizable as possible. The generalizability of LHSs depends on the chosen domain, although one should realize that the domain definition may be restricted by the characteristics of the LHSs population.[28] The broader the domain definition, given that the LHS population can appropriately represent the domain, the broader the research will be applicable and generalizable. Alternatively, evidence could be generated explicitly for internal use. Lack of generalizability of evidence generated to be used within LHSs will then be less of a problem. Yet, validity remains key.

Second, by installing safeguards to validate evidence generated within LHSs. In particular, evidence that is not published or used outside LHSs, yet the results implemented in clinical practice, should be rigorously validated. An approach could be to install independent committees comprising field experts from outside the evidence generating LHSs. This would create a system of peer review even in the absence of the traditional route of publishing. Besides performing regular peer-review of evidence they could possibly even be asked to reproduce certain results on data from outside an LHSs to falsify results in case of doubt.

Third, by seeking collaboration with other LHSs and guideline organizations from the start. To mitigate the risk of evidence conflicts between evidence created inside and outside LHSs a system should be devised on how to deal with these conflicts. For instance, conflicts could be submitted to the independent committee mentioned in the previous paragraph. Moreover, dialogue should take place on how to keep evidence compatible between different LHSs to allow meta-analytic reviews of evidence.

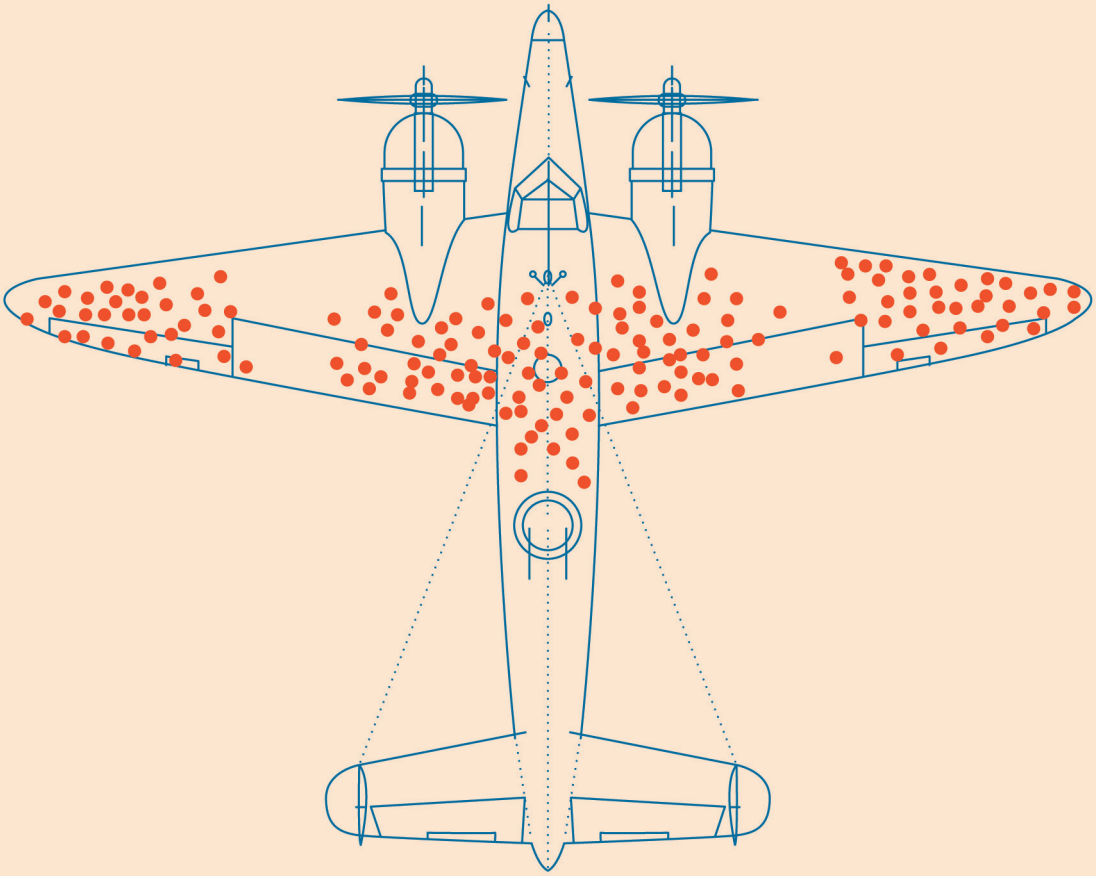
CONCLUSION

Evidence generated in LHSs will be more responsive to their populations than current knowledge and might therefore be of higher social value internally. Yet, this comes at the expense that this knowledge might have less external social value, being less applicable, precise and compatible with knowledge from outside. Moreover, this contradiction might require making choices between guideline and local recommendations in case of equal level evidence.

REFERENCES

1. Institute of Medicine. The Learning Healthcare System. Washington, DC: The National Academies Press; 2007.
2. Abernethy AP, Ahmad A, Zafar SY, Wheeler JL, Reese JB, Lyerly HK. Electronic patient-reported data capture as a foundation of rapid learning cancer care. *Med Care* 2010; 48.
3. Burnett RT, Dewanji A, Dominici F, Goldberg MS, Cohen A, Krewski D. On the relationship between time-series studies, dynamic population studies, and estimating loss of life due to short-term exposure to environmental risks. Environ. Health Perspect. 2003; 111: 1170–4.
4. Harrison MI, Shortell SM. Multi-level analysis of the learning health system: Integrating contributions from research on organizations and implementation. *Learn Heal Syst* 2020; : e10226.
5. Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? *J Am Med Informatics Assoc* 2018; 25: 963–8.
6. Maddox TM, Albert NM, Borden WB, et al. The Learning Healthcare System and Cardiovascular Care: A Scientific Statement from the American Heart Association. Circulation. 2017;134(14):e826–e857.
7. Coller BS, Califf RM. Traversing the valley of death: A guide to assessing prospects for translational success. Sci. Transl. Med. 2009; 1: 10cm9.
8. Emanual EJ, Wendler D, Grady C. What makes pediatric clinical research ethical? *JAMA* 2000; 283: 2701–11.
9. Asselbergs FW, Visseren FL, Bots ML, *et al.* Uniform data collection in routine clinical practice in cardiovascular patients for optimal care, quality control and research: The Utrecht Cardiovascular Cohort. *Eur J Prev Cardiol* 2017; 24: 840–7.
10. Habets MG, Van Delden JJ, Bredenoord AL. The social value of clinical research. BMC Med. Ethics. 2014; 15: 1–7.
11. Grady C. Ethics of international research: What does responsiveness mean? Virtual Mentor. 2006; 8: 235–40.
12. Wertheimer A. The Social Value Requirement Reconsidered. *Bioethics* 2015; 29: 301–8.
13. Wenner DM. The social value of knowledge and the responsiveness requirement for international research. Bioethics. 2017; 31: 97–104.
14. Hultcrantz M, Rind D, Akl EA, *et al.* The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol* 2017; 87: 4–13.
15. Bauer H. Ethics in Science: The Knowledge Filter. 1995. https://www.tissuegroup.chem.vt.edu/chem-ed/ethics/hbauer/hbauer-filter.html (accessed Dec 20, 2019).
16. Bauer H. Science in the 21st century: Knowledge monopolies and research cartels. *J Sci Explor* 2004; 18: 643–60.
17. Bauer H. The failings of modern medicine. *Explor J Sci Heal* 2014; 10: 345–9.
18. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet* 2017; 390: 415–23.
19. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. *J R Soc Med* 2011; 104: 510–20.
20. Riley WT, Glasgow RE, Etheredge LM, Abernethy AP. Rapid, responsive, relevant (R3) research: a call for a rapid learning health research enterprise. *Clin Transl Med* 2013; 2: 10.
21. Groenhof TKJ, Rittersma ZH, Bots ML, *et al.* A computerised decision support system for cardiovascular risk management 'live' in the electronic health record environment: development, validation and implementation—the Utrecht Cardiovascular Cohort Initiative. *Netherlands Hear J* 2019; 27: 435–42.
22. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 2015; 16: 1–14.
23. Makady A, Stegenga H, Ciaglia A, *et al.* Practical implications of using real-world evidence (RWE) in comparative effectiveness research: Learnings from IMI-GetReal. *J Comp Eff Res* 2017; 12: 485–90.

24. Kalkman S, Van Thiel G, Van Der Graaf R, *et al.* The social value of pragmatic trials. Bioethics. 2017; 31: 136–43.
25. McElwee NE, Dubois RW. From methods to policy: Enthusiasm for rapid-learning health systems exceeds the current standards for conducting it. *J Comp Eff Res* 2013; 2: 425–7.
26. Sacristán JA, Aguarón A, Avendaño-Solá C, *et al.* Patient involvement in clinical research: Why, when, and how. *Patient Prefer Adherence* 2016; 10: 631–40.
27. Grobbee DiE, Hoes AW. Clinical epidemiology: Principles, methods, and applications for clinical research. Sudbury, Mass: Jones and Bartlett Publishers, 2014.
28. Soares MO, Dumville JC, Ades AE, Welton NJ. Treatment comparisons for decision making: Facing the problems of sparse and few data. *J R Stat Soc* 2014; 177: 259–79.
29. Martina R, Jenkins D, Bujkiewicz S, Dequen P, Abrams K. The inclusion of real world evidence in clinical development planning. *Trials* 2018; 19: 698–703.
30. Prasad V, Mailankody S. Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval. *JAMA Intern Med* 2017; 177: 1569.
31. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: Promise and challenges. *Nat Rev Cardiol* 2016; 13: 350–9.
32. Britto MT, Fuller SC, Kaplan HC, *et al.* Using a network organisational architecture to support the development of Learning Healthcare Systems. *BMJ Qual Saf* 2018; 27: 937–46.
33. Prinsen CAC, Vohra S, Rose MR, *et al.* Core Outcome Measures in Effectiveness Trials (COMET) initiative: Protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set'. *Trials* 2014; 15: 247.
34. Ziman J. Public Knowledge. Cambridge: Cambridge University Press, 1968.
35. Horwitz LI, Kuznetsova M, Jones SA. Creating a learning health system through rapid-cycle, randomized testing. *N Engl J Med* 2019; 381: 1175–9.
36. Hill AB. The Environment and Disease: Association or Causation? *J R Soc Med* 1965; 58: 295–300.
37. Buccheri S, Sarno G, Fröbert O, *et al.* Assessing the Nationwide Impact of a Registry-Based Randomized Clinical Trial on Cardiovascular Practice. *Circ Cardiovasc Interv* 2019; 12: e007381.
38. Guyatt GH, Oxman AD, Vist GE, *et al.* GRADE an emerging consensus on rating quality of evidence and strength of recommendations. *Br Med J* 2008; 336: 924–6.
39. Elliott JH, Synnot A, Turner T, *et al.* Living systematic review: 1. Introduction—the why, what, when, and how. *J Clin Epidemiol* 2017; 91: 23–30.
40. Irving EA, van den Bor R, Welsing PMJ, *et al.* Series: Pragmatic trials and real world evidence: Paper 7. Safety, quality and monitoring. *J Clin Epidemiol* 2017; 91: 6–12.
41. van Dijk WB, Grobbee DE, de Vries MC, Groenwold RHH, van der Graaf R, Schuit E. A systematic breakdown of the levels of evidence supporting the European Society of Cardiology guidelines. *Eur J Prev Cardiol* 2019; 26: 1944–52.
42. Fanaroff AC, Califf RM, Windecker S, Smith Jr. S, Lopes, Renato D. Levels of Evidence Supporting American College of Cardiology/American Heart Association and European Society of Cardiology Guidelines, 2008-2018. *JAMA* 2019; 27715: 1069–80.



Chapter 6

The duty to support learning healthcare systems

Rieke van der Graaf, Wouter B van Dijk, Sara J M Laurijssen, Ewoud Schuit, Diederick E Grobbee, Martine C de Vries

American Journal of Bioethics (2020); 21 (1): 14-16.

As of October 23, 2020, almost 42 million cases of COVID-19 have been reported globally (COVID-19 Situation Update Worldwide 2020). Although many different treatments have been applied in infected people, thus far, evidence on the actual benefit of these treatments is lacking.[1] The need to rearrange the health care system in such a way as to improve the evidence on treatment for people infected with SARS-CoV-2 and to maximize the system's potential to create meaningful outcomes has been well-recognized.[2] At the same time, uncertainty and disagreement about what is best for patients is a well-known and wide-spread phenomenon in medicine, not only in times of a pandemic with a novel diseas.[3-4] In order to improve the evidence-base for people with SARS- CoV-2 infections, but also for patients with other dis- eases, Alex John London (2021) defends a Duty to Support Learning Health Systems: "*When experts dis- agree or are uncertain about the best means of pre- venting, diagnosing or treating sickness, injury or disease, medical professionals have a duty to support, and not to undermine, health systems that conduct scientifically sound and socially valuable studies in a timely manner in order to eliminate or substantially reduce this conflict or uncertainty without compromising respect for the rights and interests of study participants*.[5] Although we are sympathetic to a duty to support Learning Health Systems as defended by London, we think that the duty needs further interpretation and may be sharpened in order to be meaningful.

As we see it, London has implicitly formulated a narrow conception of this duty. First, the duty is addressed solely at medical professionals. For example, he writes that it is the medical professional who should start "scientifically sound and socially valuable research" (line 10) when there is uncertainty and disagreement about the treatment of novel conditions. Although it is not clear which experts precisely fall into this category, the literature on learning health systems informs us that a meaningful rearrangement of the health care system not only asks of medical professionals to support this system but also of others, running from boards of directors, research ethics committees, patients themselves and society at large.[6-7] Moreover, medical professionals are not supposed to work in isolation but should discuss the need and best way to solve these uncertainties and disagreements with other professionals and societal stakeholders before they start their studies. Furthermore, medical professionals need a well-functioning research infrastructure, including the funds to carry out their research, a proper data management system and sufficient and qualified personnel such as research nurses. Moreover, in order to establish a Learning Health System, research ethics committees should acknowledge the increase of research embedded in a health care system where care and research practices cannot always be easily distinguished and have a mechanism in place to review

protocols for Learning Health Systems.[6-7] Furthermore, patients and the public need to be engaged in order to create meaningful research questions and to be informed that their data may be used for large scale research projects.[6-7]

Second, from the paper as a whole it becomes clear that London seems most interested in using randomized clinical trials (RCTs) to solve states of uncertainty or disputes in health knowledge. For example, he states that "when informed medical experts are uncertain about which care is optimal, or they have definitive but conflicting preferences for different interventions, RCTs represent a way of providing access to medical interventions under conditions that support reliable inference about the relative clinical merits of those interventions".[5] Furthermore, there is no mentioning of other research methods that can be deployed in an LHS.

We agree that the use of RCTs in a pandemic and beyond is essential, in particular for obtaining evidence about novel interventions or about existing medications for novel indications. But to solve every state of uncertainty with an RCT would be virtually impossible and is not always ideal.

First, there are many types of research that do not need the RCT as a method, but still are essential for progress in a health system, for example, diagnostic or prognostic research, or the generation of hypotheses.[8]

Second, whether or not RCTs are the best means to solve disputes or states of uncertainty also depends on the type of intervention of interest. A completely new intervention may require an RCT to study its effectiveness and (short-term) safety as com- pared to care as usual; however, in the case of two interventions that have already been applied in clinical practice and where allocation of interventions is at random and not based on patient characteristics, a comparison of both interventions based on observational data may be sufficiently valid to inform clinical practice.[8]

Third, despite their methodological challenges (such as selection bias and confounding) observational studies generally have longer follow-up, allowing for assessment of long-term efficacy and safety, larger sample size, more generalizable eligibility criteria, and hard outcomes than RCT.[9] As such, evidence from observational research can 1) complement evidence from RCTs, 2) serve for hypothesis generation, and 3) serve as an intermediate basis for clinical decision making until evidence from RCTs becomes available.[10]

Fourth, the successful use of RCTs in a health system is dependent upon how well the health system itself is arranged to create meaningful and timely outcomes. For example, the so-called "Trials within Cohorts [TWiCs]" design where clinical trials are embedded in cohorts has been created with the aim to improve recruitment rates and to enable treatment comparisons.[11]

In sum, it is the organization of the health care system that needs to be transformed and let RCTs flourish optimally, provided that they are the best means to resolve knowledge conflicts. Already in 2007, the Institute of Medicine called upon health care leaders to transform their systems into learning health care systems.[12] Thus far, however, real examples of implementations of Learning Healthcare Systems have remained scarce, despite their theoretical potential.[13]

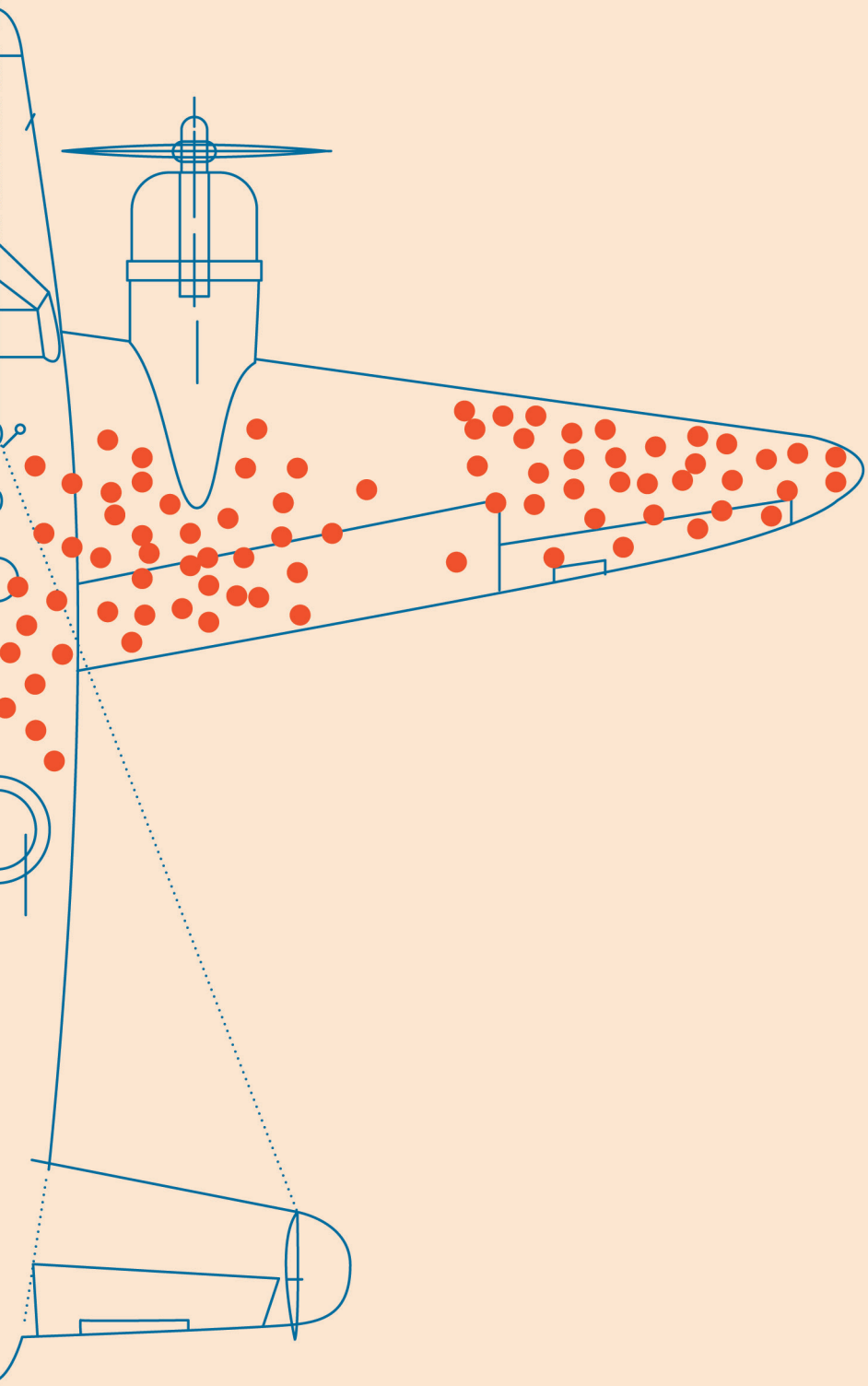
A strong Learning Health System (LHS) is essential for optimally designed and appropriately conducted RCTs. Observational data can be used as a starting point for generating hypotheses, and for a first exploration of the effectiveness and safety of interventions. Subsequently, RCTs can be used to study research questions in a more robust manner. In regard to the COVID-19 pandemic, we have seen the importance of both international collaboration and having a structure in place that is able to embed RCTs, for example, by using adaptive platform designs such as the WHO SOLIDARITY trial.[14] Currently, the SOLIDARITY Trial "is ongoing in 30 countries among the 43 countries that have approvals to begin recruiting. Overall, 116 countries in all 6 WHO regions have joined or expressed an interest in joining the trial".[14] Readiness and preparedness of health systems for such a way of learning seems essential to create progress in a pandemic. Moreover, during the outbreak of the COVID-19 pandemic (and other disease outbreaks) we have seen that time is simply lacking to afford the conduct of a robust RCT, and treatment decisions need to be taken on observational data only. The World Health Organization (WHO) emphasizes the importance of monitoring emergency use of unregistered and experimental interventions (MEURI) in a disease outbreak.[15] According to WHO, physicians overseeing MEURI have the same moral obligation to collect all scientific- ally relevant data on the safety and efficacy of the intervention as researchers who perform a clinical trial.[15] Arguably, a health system that is prepared for proper registration of characteristics and outcomes of patients treated with experimental interventions could facilitate assessment of efficacy and safety and is essential to improve the evidence base for these interventions. Eventually, in a preexistent LHS the infrastructure for data collection in this way would have already been in place.

Therefore, we defend a broad conception of the Duty to Support Learning Health Systems which is in line with the IOM's interpretation of an LHS.[12] We argue that this duty implies: *Relevant stakeholders of Health System, including community representatives, patients, boards of directors, nurses, physicians, societal stake- holders, funders, research ethics committees and researchers have a duty to rearrange their health care system in such a way that this system systematically learns from the collection, storage and use of routinely collected data in order to improve the evidence base of medicine.* The infrastructure that facilitates learning from routinely collected data may be used to embed RCTs that may help us to assess efficacy and safety of interventions but can also be used to address diagnostic and prognostic research questions by means of cohort studies. In addition, this duty includes improvements in management (infrastructure), uniform way of collecting data, patient engagement and ethical oversight systems.

In such a system, in which data are already routinely collected and studied, MEURI data can easily be merged while awaiting the conduct of robust RCTs, including adaptive platform trials, but the data can also be used to discern promising interventions from interventions that are deemed to fail. In that way, many burdensome and costly RCTs can also be prevented.

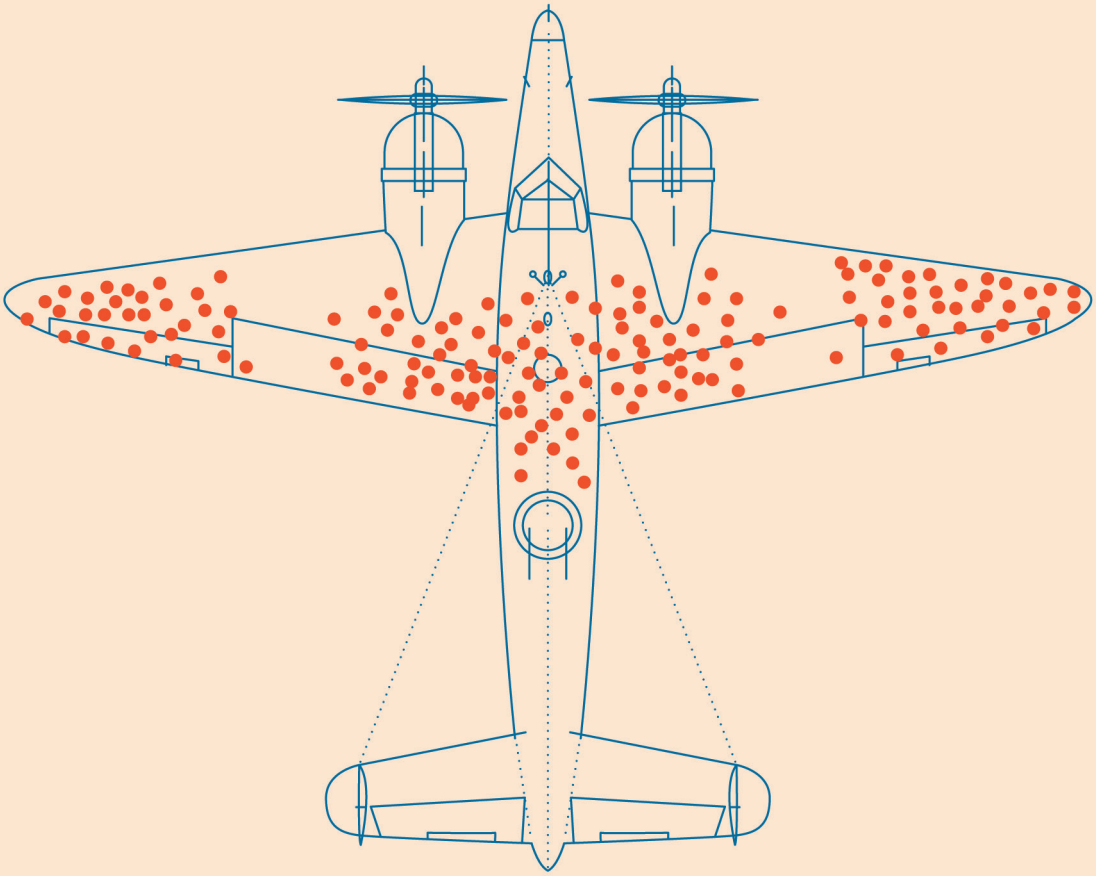
REFERENCES

1. Pan, H., R. Peto, Q. A. Karim, M. Alejandria, A. M. Henao- Restrepo, C. H. Garc?ıa, M.-P. Kieny, et al. 2020. Repurposed antiviral drugs for COVID-19–interim WHO solidarity trial results. MedRxiv
2. North, C. M., M. L. Dougan, and C. A. Sacks. 2020. Improving clinical trial enrollment — in the Covid-19 era and beyond. The New England Journal of Medicine 383 (15):1406–8.
3. Garrow, J. S. 2007. What to do about CAM: How much of orthodox medicine is evidence based? BMJ (Clinical Research ed.) 335 (7627):951.
4. Prasad, V., A. Cifu, and J. P. A. Ioannidis. 2012. Reversals of established medical practices: Evidence to abandon ship. JAMA 307 (1):37–8.
5. London, A. J. 2021. Self-defeating codes of medical ethics and how to fix them: Failures in COVID-19 Response and Beyond. The American Journal of Bioethics 21 (1): 4–13.
6. Faden, R. R., N. E. Kass, S. N. Goodman, P. Pronovost, S. Tunis, and T. L. Beauchamp. 2013. An ethics framework for a learning health care system: A departure from traditional research ethics and clinical ethics. The Hastings Center Report 43 (s1):S16–S27.
7. Wouters, R. H. P., R. van der Graaf, E. E. Voest, and A. L. Bredenoord. 2020. Learning health care systems: Highly needed but challenging. Learning Health System 4 (3): e10211.
8. Grobbee, D. E., and A. W. Hoes. 2014. Clinical epidemi- ology: Principles, methods, and applications for clinical research. Burlington, MA: Jones & Bartlett Publishers.
9. Concato, J., N. Shah, and R. I. Horwitz. 2000. Randomized, controlled trials, observational studies, and the hierarchy of research designs. New England Journal of Medicine 342:1887–92.
10. Halperin, J. L., G. N. Levine, S. M. Al-Khatib, K .K. Birtcher, B. Bozkurt, R. G. Brindis, J. E. Cigarroa, L. H. Curtis, L. A. Fleisher, F. Gentile, S. Gidding, M. A. Hlatky, J. Ikonomidis, J. Joglar, S. J. Pressler, and D. N. Wijeysundera. 2016. Further Evolution of the ACC/AHA Clinical Practice Guideline Recommendation Classification System: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Journal of the American College of Cardiology 67 (13):1572–1574.
11. Relton, C., D. Torgerson, A. O'Cathain, and J. Nicholl. 2010. Rethinking pragmatic randomised controlled trials: Introducing the "cohort multiple randomised controlled trial" design. BMJ (Clinical Research ed.) 340:C1066.
12. Olsen, L., A. Dara, and J. Michael McGinnis. 2007. Institute of Medicine (US). Roundtable on evidence-based medicine. The learning healthcare system: Workshop summary. Washington: National Academies Publishers.
13. Budrionis, A., and J. G. Bellika. 2016. The learning health- care system: Where are we now? A systematic review. Journal of Biomedical Informatics 64:87–92.
14. World Health Organization. 2020. "Solidarity" clinical trial for COVID-19 treatments. World Health Organization (WHO). Situation reports. Geneva: WHO. Accessed April 5, 2020. https://www.Who.Int/Emergencies/Diseases/ Novel-Coronavirus-2019/Global-Research-on-Novel-Coronavirus-2019-Ncov/Solidarity-Clinical-Trial-for- Covid-19-Treatments.
15. World Health Organization. 2018. Consultation on moni- tored emergency use of unregistered and investigational interventions for ebola virus disease (EVD).



PART 3

Routinely collected data in learning healthcare systems



Chapter 7

Text-mining electronic healthcare records for trial recruitment and baseline data collection

Wouter B van Dijk*, Aernoud T L Fiolet*, Ewoud Schuit, Arjan Sammani, T Katrien J Groenhof, Rieke van der Graaf, Martine C de Vriess, Marco Alings, Jeroen Schaap, Folkert W Asselbergs, Diederick E Grobbee, Rolf H H Groenwold, Arend Mosterd

Journal of Clinical Epidemiology 2021; 132: 97-105.

* Authors contributed equally

ABSTRACT

Background

Recruitment and source data collection in clinical trials is a labor intensive, expensive endeavor. Text mining of electronic healthcare records (EHRs) could reduce costs and improve efficiency through automated participant identification and baseline source data-collection. This study aimed to validate participant recruitment and compare extracted baseline data using EHR text mining of structured and unstructured data to those from an ongoing trial on the effects of low dose colchicine in patients with stable coronary artery disease.

Methods

In three medical centers with different EHR vendors, EHR-based text-mining was used to automatically screen patients for trial eligibility and extract baseline data on nineteen characteristics. First. the yield of screening with automated EHR text-mining search was compared with manual screening by research personnel. Second, accuracy of extracted baseline data by EHR text mining was compared to manual data entry by research personnel

Results

568 (0.6%) of 92,466 patients visiting the out-patient cardiology departments were enrolled in the trial during its recruitment period using manual screening methods. Automated EHR data screening of all patients showed that the number of patients needed to screen could be reduced by 73,863 (79.9%). The remaining 18,603 (20.1%) contained 458 of the actual participants (82.4% of participants).

In trial participants, automated EHR text-mining missed a median of 2.8% (Interquartile range [IQR] across all variables 0.4-8.5%) of all data points compared to manually collected data. Overall accuracy of automatically extracted data was 88.0% (IQR 84.7-92.8%).

Conclusions

Automatically extracting data from EHRs using text-mining can be used to identify trial participants and to collect baseline information.

INTRODUCTION

Clinical research requires highly detailed information on large numbers of subjects, often acquired by many investigators and supporting staff. In particular, prospective research such as registries and randomized clinical trials (RCT) need to comply with high standards of data validity.[1,2] Scientific and regulatory requirements make such endeavors laborious and increase costs to a level only large companies are able to meet.

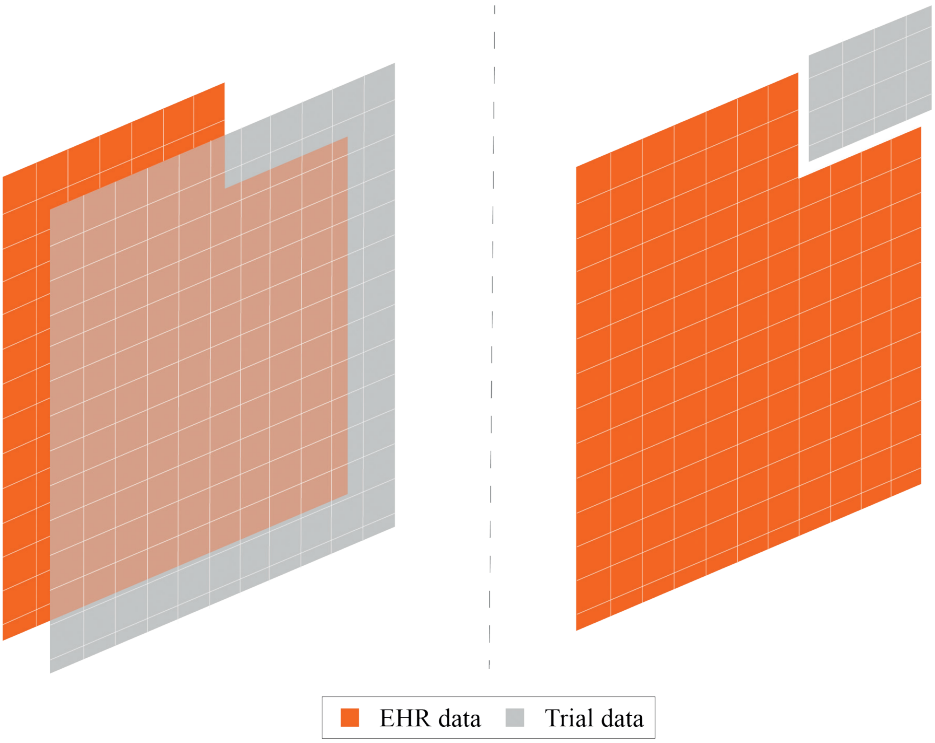
Cardiovascular outcome trials with moderate to low absolute risks nowadays require over 10,000 participants and are estimated to cost between 35,000 and 45,000 US dollars per participant, with total costs for conduct up to half a billion US dollars.[3,4] A major part of these costs are attributable to participant recruitment and follow-up, for a large part comprising data-collection.[5,6] Standing practice for clinical trials is that dedicated personnel enters source data in distinct (electronic) clinical report forms (CRFs). This data, however, is generally already collected in clinical care and available in electronic healthcare records (EHRs), thus creating overlapping copies of data that are already available (Figure 1a).

Automated EHR data-mining may provide a valuable method to complement or even substitute current data-collection methods,[7] which could save up to one third of recruitment costs.[8] In recent years, several supervised patient-diagnosis registries with labeled clinical data emerged to improve trial efficiency.[9] The use of automatically collected EHR data in trials, however, is still very limited.[10] Conventional data-collection methods generally involve retrieving information through researcher-patient interviews and manual data extraction. After retrieval, data is then entered manually in electronic data capture (EDC) systems, as part of CRFs. Data quality is guaranteed up to a certain level by automated control processes, and internal and external monitoring.[11] If EHR data are to be used to identify participants or as alternative data source, these data should be of sufficient quality. High data quality is paramount, yet will differ per objective. The accuracy level is relative to the nature of data. Outcome data that is used to estimate a treatment effect requires higher fidelity than baseline data.[12]

We hypothesized that patients eligible for trial participation can be effectively identified on information already present in EHRs using automated text-mining. Second, we hypothesized that the majority of data collected for the purpose of the trial is also already available in EHRs. If extracted automatically with acceptable accuracy, the extensive

manual entry by investigators in EDCs could be reduced . If true, data collection efforts could focus on information not available from EHRs and reduce manual EHR-to-EDC data duplication that now is common (**Figure 1**).

Figure 1: Layers of data collected during trials (left: required trial data collection when not using EHR data in perspective to data available in EHR; right: (theoretical) required trial data collection when using EHR data)



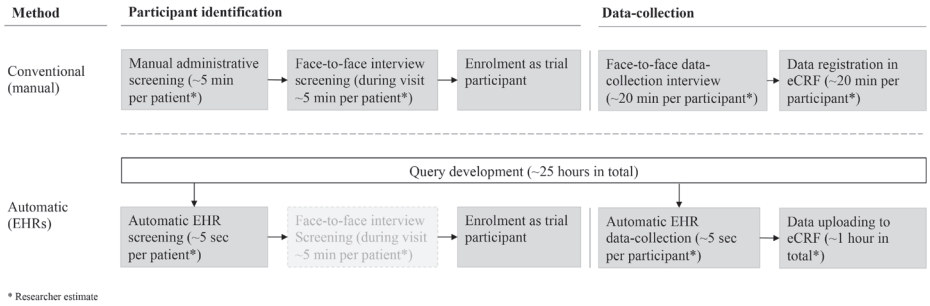
METHODS

This study was a multicenter, multi-EHR-vendor, validation study to assess accuracy of automated EHR text-mining for trial participant screening and baseline data-collection. As a reference standard, we used manual participant screening and data collection by manual data entry in EDCs, which is the current standard for most RCTs.

First, all patients who visited the out-patient cardiology clinics of the three participating medical centers during the recruitment phase (October 1, 2016 – December 1, 2018) of the LoDoCo2 trial were automatically and anonymously screened retrospectively for

eligibility of participation in the trial according to its inclusion and exclusion criteria (**Figure 2**). The yield of eligible patients via this method was compared to those actually included in the trial by manual screening for trial participation. Second, baseline characteristics were automatically collected for all trial participants, and accuracy was assessed against manually collected data.

Figure 2: Overview of the process of conventional and automated participant identification and data collection, and associated estimated time of these processes



The LoDoCo2 trial

Conventional participant identification and data-collection methods used in the international clinical trial LoDoCo2 were used as reference standard. The LoDoCo2 trial was chosen as it represents a prototype large international multicenter cardiovascular outcome trial.

In short, the LoDoCo2 trial was a randomized investigator initiated international, multicenter study that investigated whether colchicine 0.5 mg once daily as compared to placebo in patients with stable coronary artery disease reduces the incidence of major adverse cardiovascular events.[13] The trial's recruitment started in December 2016 and was completed in December 2018. The trial methodology and results have been reported before.[14]

Study population

This study was based on the data of patients visiting the cardiology out-patient clinics of three large Dutch medical centers. The medical centers were selected to represent the major EHR software vendors in the Netherlands (Epic [Hospital A], Chipsoft [Hospital B], CSC Care solutions [Hospital C]; cumulatively used in 80% of the Dutch hospitals and almost 10% of the hospitals world-wide[15,16]).

Participants of the LoDoCo2 trial were retrieved on their trial identification number and unique on-site identifier as recorded in their EHR files. Participants for which no trial identifiers were reported in the EHR were ignored, since they could not be linked to CRF data functioning as reference standard.

Participant identification methods

Automatic, using text-mining from EHRs

A Boolean retrieval query to obtain required data was developed in adherence with the eligibility criteria of the LoDoCo2 trial by two authors (WBvD and ATLF) (**Supplemental materials, Table S1a**). To develop the query a graphic user interface data mining tool with text-mining features was used (CTcue, version 2.0.12; Amsterdam, The Netherlands). This data mining tool integrally searched structured and unstructured EHR data (including clinical letters, in-hospital consultations, procedures, diagnostic tests and drug prescriptions).

Both authors who developed the query were considered to have content expertise from their medical backgrounds and had extensive experience in query development. Additionally, one of these authors (ATLF) was also a lead investigator of the LoDoCo2 trial.

The query consisted of regular expressions of the eligibility criteria as given by the LoDoCo2 trial, their synonyms and negations (e.g., 'no hypertension' instead of 'hypertension'). Synonyms were added using the automatic synonym expander built into the data mining tool and supplemented with synonyms and abbreviations commonly used by the query developing authors (**Supplemental materials, Table S1a**).

To preclude automatic retrieval of information entered in the EHR after trial participation, only data registered in EHRs prior to the screening of the trial were used. No site-specific optimizations were added to the query, except for the retrieval of trial participants and peri-procedural drug recognition adjustments. To approximate data-collection as would have been performed in the trial, the most recent status on any data point before entering the trial was taken. Additionally, drug use data were limited to data registered within a year of enrollment. When no measurement of a variable was found it was assumed to be absent for the participant.

Manual participant identification, as used in the LoDoCo2 (reference standard)
Trial investigators of the LoDoCo2 trial used two steps to identify trial participants. First, manual screening was performed for eligibility using the EHR files prior to their outpatient clinic visit. Second, patients were interviewed face-to-face to verify eligibility and ask for participation. After providing informed consent, participation in the trial ensued.

Baseline data extraction methods
Automatic, using text-mining from EHRs
A query was developed to automatically collect data form the EHRs on nineteen variables, which contained information about demography, medical history, procedure history and drug use as reported in the baseline table of the trials' methods paper (Supplement 1b) For development of this query the same methods were employed as for the participant identification query.

Conventional data extraction, as used in the LoDoCo2 trial (reference stanard)
In the LoDoCo2 trial data were collected manually during face-to-face baseline interviews at trial enrolment with participant. Interview data was first recorded as source data on-site and afterwards entered in the trial's EDC system.

Analysis
Participant identification efficiency
For each site the number of unique patient visits during the trial recruitment period, number of patients automatically identified as potentially eligible and number of patients enrolled in the trial were recorded and compared to the number of patients enrolled in the actual trial. For both methods a theoretical yield was calculated based on the patients needed to screen for identification. To determine the yield of the automatic participant identification the number of enrolled trial participants was used as a proxy since it was not possible to assess how many of the automatically identified potentially eligible patients would have been enrolled retrospectively.

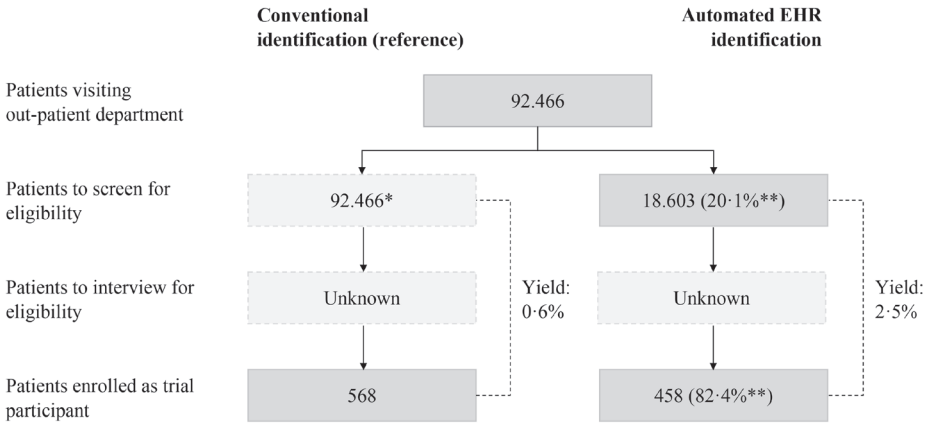
Data-collection accuracy
Results of automated EHR text-mining were compared to manually collected trial data on their distributions and accuracy (defined as: [true positive data points + true negative data points] / all data points) on an individual patient level. For clarity and to show agreement between EHR vendors, accuracies of the various medical centers were plotted against the overall accuracy in a forest plot.

RESULTS

Participant identification efficiency

A total of 92,466 patients visited the cardiology out-patient clinic of the three study centers during the recruitment period of the LoDoCo2 trial (October 1, 2016-December 1, 2018). Of these, 568 patients (0.6%) were enrolled in the LoDoCo2 trial. (**Figure 3, Table 1**).

Figure 3: Eligible patients identified with conventional and automated participant identification



* Number of patients theoretically screened by trial investigators.
** Percentages are calculated in relation to the number of patients using conventional identification.

Table 1: Number of patients visiting, eligible and enrolled per participating hospital

Hospital	Enrolment period	Total no. of patients visiting	No. of trial participants (%)	No. of visiting patients potentially elligible (%)	No. of trial participants identified (%; % of participants)
A	Feb 2017 – Oct 2018	51,943	169 (0.3)	10,705 (20.6)	151(1.4; 89.3)
B	Jul 2017 – Oct 2018	14,206	69 (0.5)	2,966 (20.9)	65 (2.2; 94.2)
C	Oct 2016 – Dec 2018	26,317	330 (1.3)	4,932 (18.7)	252 (5.1; 76.4)
Total		92,466	568 (0.7)	18,603 (20.1)	468 (2.5; 82.4)

For the LoDoCo2 trial, all patients visiting the cardiology out-patient clinics were screened on trial eligibility. Automated EHR data screening resulted in a reduction of 73,863 (79.9%) patients that needed to be screened for trial participation. The remaining 18,603 (20.1%) contained 458 of the actual trial participants (82.4% of participants). Further inspection of the 110 (17.6%) trial participants missed by the data mining tool showed that in the

automatically retrieved data on one or more inclusion or exclusion criteria were missing (no proof of coronary artery disease [found as a coronary angiography; CT coronary angiography or Coronary Artery Calcium Score]: n=38; no known renal function: n=41; date of previous Coronary Artery Bypass unknown: n=41). Characteristics of missed participants did not differ substantially from identified participants (median difference of all variables 1.6%, IQR 3.1%), values were therefore assumed to be missing at random.

Data-collection accuracy

Of the 568 trial participants, 540 (95.1%) enrolled trial participants were automatically retrieved on their trial identification number or unique on-site identifier with the data mining tool.

On aggregate level, availability of baseline characteristics for participants using automated EHR text-mining differed by 2.8% (median; IQR across all variables 0.4-8.5%) with manually collected trial data (Table 2; center-specific distributions are presented in supplement 2a). Notably larger differences between automated EHR text-mining data and manually collected trial data were found for hypertension (26.2%), antiplatelet therapy (29.1%) and beta-blocker use (24.4%).

On an individual participant level, automated EHR text-mining data showed 88.0% accuracy (median; IQR 84.7-92.8%) when compared to the conventionally collected trial (**Table 2**; center-specific accuracy is presented in **Supplemental materials, Table S2b**). Overall, 9.8% of the data extracted from EHRs were false positive (i.e., data on a variable present in EHR data and not present in trial data), and 3.1% false negative (i.e., data on a variable not present in EHR data and present in trial data) (**Table 3**; for contingency tables of different medical centers see **Supplemental materials, Table S2c**). Of all data points, positive predictive value was 0,928, negative predictive value was 0,937, sensitivity was 0,806, specificity was 0,827 and F1-score was 0,863 (for test performance scores of individual variables see **Supplemental materials, Table S2d**). Lowest accuracies were found for hypertension (62.6%), antiplatelet therapy (68.8%) and beta blocker use (73.3%). Accuracies for hypertension, antiplatelet therapy and beta blocker therapy differed between the participating medical centers, with hypertension ranging from 52.2% to 64.2%, antiplatelet therapy from 60.3% to 86.4% and beta blocker use ranging from 66.4% to 84.7%.

Table 2: Distributions and accuracy of baseline variables automatically collected from EHR data compared to trial data

Variable	Trial data, %	EHR data, %	Absolute difference, %	Agreement, %
Sex (Male)	83.6	82.8	-0.8	99.6
Current smoker (Yes)	14.2	13.5	-0.7	85.2
Demographics, median (IQR)			*-0.7 (-0.8--0.7)*	*92.4 (88.8-96)*
Hypertension (Yes)	53.6	79.8	26.2	62.6
Diabetes (Yes)	18.4	20.4	2.0	95.0
Insulin dependent diabetes (Yes)	5.8	13.3	7.5	90.8
Renal function (Not impaired)	91.8	88.5	-3.3	91.4
Prior ACS (Yes)	79.6	84.8	5.2	84.2
Prior PCI (Yes)	87.0	91.9	4.9	92.8
Prior CABG (Yes)	11.5	12.0	0.5	97.6
Atrial fibrilation (Yes)	13.7	8.0	-5.7	86.4
Gout (Yes)	7.3	7.6	0.3	92.8
Medical history, median (IQR)			*2 (0.3-5.2)*	*91.4 (86.4-92.8)*
Antiplatelet therapy (APT) (Yes)	69.2	98.3	29.1	68.8
Oral anticoagulant therapy (OAC) (Yes)	14.8	22.8	8.0	91.8
No APT or OAC (Yes)	0.4	1.1	0.7	98.8
Statin (Yes)	91.8	93.1	1.3	87.2
Ezetimibe (Yes)	23.5	26.3	2.8	87.0
ACE Inhibitor (Yes)	70.3	79.3	9.0	88.0
Beta blocker (Yes)	67.5	91.9	24.4	73.6
Calcium channel blocker (Yes)	27.7	41.5	13.8	80.8
Drug use, median (IQR)			*8.5 (2.4-16.5)*	*87.1 (79-89)*
Overall, median (IQR)			***2.8 (0.4-8.5)***	***88 (84.7-92.8)***

△ Hospital A, + Hospital B, × Hospital C

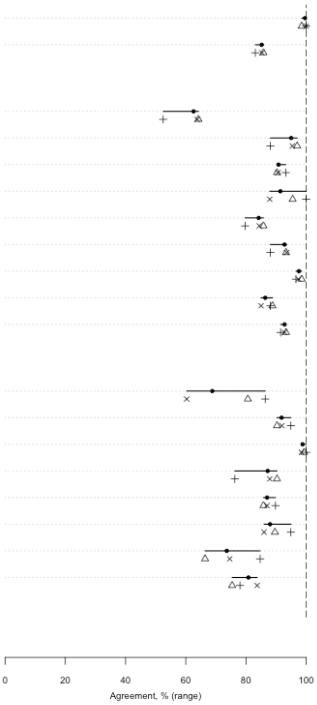


Table 3: Overall contingency table of the accuracy of collected baseline variables

		Trial data, no (%)		
		True***	False****	***Overall***
Automatically collected EHR data	True*	3,855 (40.6)	929 (9.8)	***4,784 (53.4)***
	False**	299 (3.1)	4,417 (46.5)	***4,716 (49.6)***
	Overall	***4,154 (43.7)***	***5346 (56.3)***	***9,500 (100)***

* Data on a variable present in EHR data; ** Data on a variable not present in EHR; *** Data on a variable present in trial data; **** Data on a variable not present in trial data

DISCUSSION

This study shows that it is feasible to use automated EHR text-mining to identify eligible trial participants and collect baseline data. By identifying eligible patients, only 20.1% of the original 92,466 visiting patients had to be screened manually for trial inclusion. In

these 20.1%, 82.4% of the participants were present. Data extracted from EHRs showed an average accuracy of 87.1% to the manually collected data of the LoDoCo2 trial.

Several studies have investigated the opportunities of using EHRs for recruitment and data-collection in clinical research and trials but only few compare EHR data to trial data.[17–21] In general, studies focusing on assessing EHR data quality showed mixed results.[10,22–24] Results from studies focusing on structured EHR data and text-mining in separate EHR components generally showed low yields for EHR quality data.[22–24] A study from 2013 assessed the completeness of structured EHR data to trial eligibility criteria originating from multiple trials, showing that 35% of the patient characteristics derived from the eligibility criteria were available in structured EHR data at the time.[23] In the same year, EHR medication lists were shown to have a very broad accuracy (10-90%).[22] Studies automatically text-mining EHRs integrally, however, reported more favorable results with accuracies comparable to those found in this study.[10,24] In addition, registries based on routinely collected data have been reported to be of high value for trial recruitment and data-collection.[25]

Implications for using EHR data in clinical research

When quality of EHR data extraction is of an acceptable level it could improve efficacy in trial conduct. As such, EHR data-collection would allow reallocation of resources and a reduction in execution costs.[7]

Participant identification efficiency

Using automated EHR text-mining, we were able to identify patients potentially eligible for trial participation. These results are in line with results found by previous studies.[18,19,26] In participant recruitment, a high positive predictive value using automated EHR participant screening (i.e. most patients screened as positive also enroll in the trial) would maximize efficacy improvements.[27] Our study indicates that automated EHR screening has the potential to identify large numbers of eligible participants in a time- and cost efficient manner (data not shown).

Data-collection accuracy

Since baseline characteristics are not always included in final outcome analysis generally, small errors in these data can be acceptable when counterbalanced by improved efficiency. Incorporation of baseline characteristics measured with error in the analyses would only have an effect on research validity when accuracy is not randomly

distributed across intervention groups. If random, it could affect precision of effect estimates after adjustment.[12]

Accuracy of automated EHR data-collection depends on the amount of missing data and measurement errors. First, variables collected from data can be missing because they were not recorded or not extracted from the data. Physicians often measure and register only what they consider relevant for delivering care. Consequently, (ordinary) characteristics that are desired in clinical research are not registered.[28] Whether this will lead to problems in identifying patients eligible for trial participation differs per variable and context. Missing data on smoking for example will be of less value than missing data on coronary revascularization since clinicians will not always ask about smoking but may be expected to document coronary interventions.[29] These factors make it harder to extract data due to ensuing variability in how characteristics are reported. Substantive knowledge on the topics of data to be extracted is therefore essential still.

Second, EHR data could contain more measurement errors because they were not collected and measured in a standardized format as is generally done in conventional trial data collection. EHR data can, for example, be hampered in its currency (i.e. stored variables are out of date) due to irregular visits of patients. These remain challenges of the use of EHR data that should be addressed in future research.

Third, relevant information encompassed in the EHR can still be missed due to interindividual differences in reporting or reporting errors (abbreviations, misspelling, synonyms). Improved intelligent text-pattern recognition systems might reduce the risk for missing data.

Future perspectives

EHR data-collection will probably be best used in conjunction with other data-collection methods instead of replacing them. In the design of trials, investigators can take automated and manual EHR data-collection in account in the design phase of the trial. Our results show that automated EHR screening for eligible patients might result in a somewhat different study population compared to population currently enrolled. Effects on generalizability should be considered, although the resulting patient population might well reflect a more real world sample of participants if their characteristics differ from the original study population.[18] Benefits of increased efficiency in the identification of eligible

patients might make it easier to enroll patients and as such reach the desired number of inclusion faster than with conventional participant recruitment.

Study limitations

This study combined data from multiple medical centers, all using different EHR software vendors, and shows consistent results for the broad range of systems. Yet, three main limitations should be noted on it.

First, accuracy of information on hypertension, antiplatelet therapy and beta blockers deviated notably from collected trial data. Deviation between EHR and trial data was probably due to hypertension being defined as "using antihypertensive drugs" in the LoDoCo2 trial, which was hard to mirror in the EHR search query. Deviations on drug prescriptions and use variables were mainly attributed to registered timeframes of drugs and insufficient indexing of hospital drug prescription systems by the data extraction tool. Moreover, hospital physicians might not have registered home prescriptions for all patients adequately deviating results on drugs too.

Second, it was assumed that all patients visiting the out-patient cardiology clinics of the three hospitals were screened conventionally for participation in the LoDoCo2 trial. If this was not the case expected yield of automated participant identification would be overestimated in this study.

Third, our Boolean query was not enhanced with natural language processing algorithms because of limitations of the employed data mining tool and language specific limitations. Text-mining was therefore interpreted broadly as the ability to automatically extract information from unstructured texts.

CONCLUSIONS

Data extracted from EHRs using text-mining can be used to identify patients eligible for trial participation and for the collection of baseline characteristics. This method might substantially reduce time and costs related to recruitment and data-collection in clinical trials. Whether this premise can be realized depends on whether small accuracy losses are deemed acceptable in the context of the trial that is performed. This study focused on patient eligibility screening and participant baseline data-collection; future research is needed to assess the quality of outcome detain EHRs.

REFERENCES

1. The European Medicines Agency Working Group on Clinical Trials conducted outside of the EU/EEA. Reflection Paper on Ethical and GCP Aspects of Clinical Trials of Medicinal Products for Human Use Conducted Outside of the EU/EEA and Submitted in Marketing Au. Vol 44.; 2012.
2. U.S. Food and Drug Administration. Use of Electronic Health Record Data in Clinical Investigations Guidance for Industry U.; 2019.
3. Moore TJ, Zhang H, Anderson G, Alexander GC. Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015-2016. JAMA Intern Med. 2018;178(11):1451.
4. Solomon SD, Pfeffer MA. The Future of Clinical Trials in Cardiovascular Medicine. Circulation. 2016;133(25):2662-2670.
5. Bentley C, Cressman S, van der Hoek K, Arts K, Dancey J, Peacock S. Conducting clinical trials—costs, impacts, and the value of clinical trials networks: A scoping review. Clin Trials. 2019;16(2):183-193.
6. Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? Nat Rev Drug Discov. 2017;16(6):381-382.
7. McClellan M, Brown N, Califf RM, Warner JJ. Call to Action: Urgent Challenges in Cardiovascular Disease: A Presidential Advisory From the American Heart Association. Circulation. 2019:1-11.
8. Sertkaya A, Birkenbach A, Berlind A, Eyraud J. Examination of Clinical Trial Costs and Barriers for Drug Development; 2014.
9. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform. 2017;26(1):38-52.
10. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: A systematic review. J Am Med Informatics Assoc. 2016;23(5):1007-1015.
11. Vantongelen K, Rotmensz N, Van Der Schueren E. Quality control of validity of data collected in clinical trials. Eur J Cancer Clin Oncol. 1989;25(8):1241-1247.
12. Chan SF, Macaskill P, Irwig L, Walter SD. Adjustment for baseline measurement error in randomized controlled trials induces bias. Control Clin Trials. 2004;25(4):408-416.
13. Nidorf SM, Fiolet ATL, Eikelboom JW, et al. The effect of low-dose colchicine in patients with stable coronary artery disease: The LoDoCo2 trial rationale, design, and baseline characteristics. Am Heart J. 2019;218:46-56.
14. Nidorf SM, Fiolet ATL, Mosterd A, et al. Colchicine in Patients with Chronic Coronary Disease. N Engl J Med. 2020:1-10.
15. KPMG. EHR Vendor Market. 2017.
16. Zorgvisie. Het complete epd-overzicht: welk ziekenhuis heeft welke leverancier? Houten: Bohn Stafleu van Loghum 2019.
17. Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: new opportunities for clinical research. J Intern Med. 2013;274(6):547-560.
18. Lai YS, Afseth JD. A review of the impact of utilising electronic medical records for clinical research recruitment. Clin Trials. 2019;16(2):194-203.
19. Schreiweis B, Trinczek B, Köpcke F, et al. Comparison of Electronic Health Record System Functionalities to support the patient recruitment process in clinical trials. Int J Med Inform. 2014;83(11):860-868.
20. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol. 2017;106(1):1-9.
21. Tissot H, Shah A, Agbakoba R, et al. Natural Language Processing for Mimicking Clinical Trial Recruitment in Critical Care: A Semi-automated Simulation Based on the LeoPARDS Trial. medRxiv. September 2019:19005603.
22. Walsh KE, Marsolo KA, Davis C, et al. Accuracy of the medication list in the electronic health record—implications for care, research, and improvement. J Am Med Informatics Assoc. 2018;25(7):909-912.

23. Köpcke F, Trinczek B, Majeed RW, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Med Inform Decis Mak. 2013;13(1):37.
24. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. BMC Med Inform Decis Mak. 2006;6:1-9.
25. Buccheri S, Sarno G, Fröbert O, et al. Assessing the Nationwide Impact of a Registry-Based Randomized Clinical Trial on Cardiovascular Practice. Circ Cardiovasc Interv. 2019;12(3):e007381.
26. Sumi E, Teramukai S, Yamamoto K, Satoh M, Yamanaka K, Yokode M. The correlation between the number of eligible patients in routine clinical practice and the low recruitment level in clinical trials: a retrospective study using electronic medical records. Trials. 2013;14(1):426.
27. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol. 2012;65(3):343-349.e2.
28. Kruse CS, Stein A, Thomas H, Kaur H. The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature. J Med Syst. 2018;42(11):214.
29. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform. 2013;46(5):830-836.

SUPPLEMENTARY MATERIALS

Table S1a: Search query used to identify potential eligible patients

Type	Criterium	Search terms
All	Appointment cardiology outpatient department during recruitment periode	Appointment cardiology department between [START_RECRUITMENT] and [END_RECRUITMENT]
Inclusion criteria	1. Are aged >35 and <82 years,	Age >35 and <82 years
	2. Have proven coronary artery disease; as evidenced by coronary angiography, CT coronary angiography or a Coronary Artery Calcium Score (Agatston score >400). Individuals with a history of bypass surgery are only eligible if they have undergone coronary artery bypass surgery more than 10 years before or have angiographic evidence of graft failure or have undergone percutaneous intervention since their bypass surgery.	"Percutane Coronaire Interventie" OR "Percutaneous Coronary Intervention" OR "Percutaneous Coronary Revascularization" OR "pci" OR "bms" OR "bvs" OR "des" OR "stent" OR "dotter" OR "CT coronary angiography" OR "Agatston" OR (("CABG" OR "Coronary artery bypass")
	3. Have been clinically stable for at least six months.	AND NOT < 6 months
Exclusion criteria	1. Women who are pregnant, breast feeding or may be considering pregnancy during the study period,	NOT ("zwangerschap" OR "zwanger" OR "borstvoeding")
	2. Have renal impairment as evidenced by a serum creatinine >150 µmol /l or estimated glomerular filtration rate (eGFR) <50mL/min/1.73m2,	MEASUREMENT eGFR <50 OR serum creatinine >150
	3. Have severe heart failure – systolic or diastolic New York Heart Association Functional classification 3 or 4,	NOT ("NYHA III" OR "NYHA IV")
	4. Have moderate or severe valvular heart disease considered likely to require intervention,	NOT ("valvular heart disease OR "hartklep ziekten")
	5. Are dependent or frail or have a life expectancy < five years,	NOT ("frail" OR "beperkte levensverwachting")
	6. Have peripheral neuritis, myositis or marked myo- sensitivity to statins,	NOT ("neuritis" OR "myositis" OR "myo-sensitivitis"
	7. Are already taking long term colchicine therapy for any other reason,	NOT DRUG "Colchicine"
	8. Are currently enrolled in a competing trial.	NOT [RUNNING TRIALS]

Table S1b: Search query used to automatically collect variables representing baseline characteristics from EHR data

Variable	Search terms
LoDoDoCo2 first mention	[Hospital A] “Betreft: deelname LoDoCo2 onderzoek” [Hospital B] “Betreft: deelname LoDoCo2 onderzoek” [Hospital C] DRUG “LoDoCo2 studiemedicatie (NIET run-in)”
LoDoCo2 trial identifier	[Hospital A] “L1004” [Hospital B] “L1007” [Hospital C] “L1002”
Current smoker	[ANSWER: No]”roken nee” OR “rookt niet” OR “roken gestaakt” OR “roken niet” OR “gestopt met roken” OR “roken gestopt” OR “ex roker” OR “ex-roker” [ANSWER: Yes] “patient rookt” OR “roken wel” OR “rookt wel” OR “roken ja” OR “rookt sinds” OR “rookt ja” OR “rookt” OR “roker”
Hypertension	[ANSWER: No] “Geen hypertensie” OR “hypertensie geen” OR “hypertensie nee” OR “geen hoge bloeddruk” OR “hoge bloeddruk nee” OR “HT nee” OR “geen HT” OR “geen hyperRR” OR “hyperRR nee” [ANSWER: Yes] “Hypertensie ja” OR “wel hypertensie” OR “hoge bloeddruk ja” OR “wel hoge bloeddruk” OR “bekend met hypertensie” OR “hyperRR ja” OR “bekend met hyper RR” OR “bekend met hoge bloeddruk” OR “HT ja” OR “wel HT” OR “bekend met HT”
Diabetes	[ANSWER: No] “Geen diabetes” OR “Diabetes nee” OR “DM nee” OR “DM type 1 nee” OR “DM type I nee” OR “DM type 2 nee” OR “DM type II nee” [ANSWER: Yes] “DMI” OR “DM1” OR “Dmt1” OR “DM I” OR “IDDM” OR “IADM” OR “Type I Diabetes” OR “Diabetes Type I” OR “Juveniele Diabetes” OR “Type I Diabetes Mellitus” OR “Type 1 Diabetes Mellitus” OR “Diabetes Mellitus Type 1” OR “Diabetes Mellitus Brittle” OR “Insuline-afhankelijke Diabeet” OR “Insuline-afhankelijk Diabetes” OR “Diabetes Insuline-afhankelijk” OR “Insulin-dependent Diabetes Mellitus” OR “Diabetes Mellitus Insuline-afhankelijk” OR “Insuline-afhankelijke Diabetes Mellitus” OR “Diabetes Mellitus Insuline-afhankelijke” OR “Mellitus Insuline-afhankelijke Diabetes” OR “DMI” OR “DM1” OR “DM I” OR “IDDM” OR “IADM” OR “Type 1 Diabetes” OR “Juvenile Diabetes” OR “Diabetes Mellitus Type I”, OR “Type 1 Diabetes Mellitus” OR “Diabetes Mellitus, Type 1” OR “DM type 1” OR “DM type OR “DM2” OR “Dmt2” OR “DMII” OR “MODY” OR “Niddm” OR “DM II” OR “NIADM” OR “Diabetes Niddm” OR “Niddm Diabetes” OR “Ouderdomssuiker” OR “Diabetes Type II” OR “Type II Diabetes” OR “Ouderdomsdiabetes” OR “Ouderdoms Diabetes” OR “Maturity-onset Diabetes” OR “Diabetes Maturity-onset” OR “Diabetes Mellitus Type 2” OR “Type 2 Diabetes Mellitus” OR “Type II Diabetes Mellitus” OR “Diabetes Mellitus Stabiele” OR “Niet-insuline-afhankelijk Diabetes” OR “Diabetes Niet-insuline-afhankelijk” OR “Maturity-onset Diabetes of The Young” OR “Diabetes Mellitus Insuline-onafhankelijk” OR “Niet-insulineafhankelijke Diabetes Mellitus” OR “Niet-insuline-afhankelijke Diabetes Mellitus” OR “DM2” OR “DMII” OR “MODY” OR “DM II” OR “NIDDM” OR “NIADM” OR “Type 2 Diabetes” OR “Maturity-onset Diabetes” OR “Type 2 Diabetes Mellitus” OR “Diabetes Mellitus Type II” OR “Diabetes Mellitus, Type 2” OR “Diabetes Mellitus Non-insulin-dependent” OR “T2dm - Type 2 Diabetes Mellitus” OR “DM Type 2” OR “DM type II” I” OR”Diabetes ja” OR “wel diabetes” OR “bekend met diabetes” OR “DM ja” OR “wel DM” OR “bekend met DM” OR “DM type II ja” OR “DM type 2 ja” OR “diabetes mellitus ja” OR “wel diabetes mellitus” OR “diabetes mellitus type II ja” OR “diabetes mellitus type 2 ja”

Table S1b: *(continued)*

Variable	Search terms
Insulin dependent diabetes	[ANSWER: No] "Geen IDDM" OR "Geen insuline afhankelijkheid" OR "Niet afhankelijk van insuline" OR "IDDM nee" [ANSWER: Yes] (DRUG "Insulin" OR "Insulatard" OR "Actarapid" OR "Novorapid") OR ("DMI" OR "DM1" OR "Dmt1" OR "DM I" OR "IDDM" OR "IADM" OR "Type I Diabetes" OR "Diabetes Type I" OR "Juveniele Diabetes" OR "Diabetes Juvenile-onset" OR "Juvenile-onset Diabetes" OR "Type I Diabetes Mellitus" OR "Type 1 Diabetes Mellitus" OR "Diabetes Mellitus Type 1" OR "Diabetes Mellitus Brittle" OR "Diabetes Mellitus Juveniel" OR "Diabetes Mellitus Juveniele" OR "Insuline-afhankelijke Diabeet" OR "Insuline-afhankelijk Diabetes" OR "Diabetes Insuline-afhankelijk" OR "Diabetes Mellitus 'sudden-onset'" OR "Diabetes Mellitus Ketose-gevoelige" OR "Insulin-dependent Diabetes Mellitus" OR "Diabetes Mellitus Insuline-afhankelijk" OR "Insuline-afhankelijke Diabetes Mellitus" OR "Diabetes Mellitus Insuline-afhankelijke" OR "Mellitus Insuline-afhankelijke Diabetes")
Renal impairment	MEASUREMENT eGFR <50 OR serum creatinine >150
Prior ACS	[ANSWER: Yes] "ACS" OR "Acuut Coronairsyndroom" OR "ACS" OR "Acute Coronary Syndrome" OR "acuut coronair syndroom" OR "STEMI" OR "Nstemi" OR "N-stemi" OR "Non-stemi" OR "Hartaanval" OR "Hartinfarct" OR "Ipl-infarct" OR "Hart Aanval" OR "Hart Infarct" OR "Hartaanvallen" OR "Hartinfarcten" OR "Myocardinfarct" OR "Hart Infarcten" OR "Septaal Infarct" OR "Infarct Myocard" OR "Voorwandinfarct" OR "Myocard Infarct" OR "Voorwand Infarct" OR "Cardiaal Infarct" OR "Posteriorinfarct" OR "Myocardinfarcten" OR "Achterwandinfarct" OR "Onderwand Infarct" OR "Myocard Infarcten" OR "Achterwand Infarct" OR "Anterolateraal Infarct" OR "Subendocardiaal Infarct" OR "Infero-lateraal Infarct" OR "Infero-posterior Infarct" OR "Infero-postero-lateraal Infarct" OR "Rust Angina" OR "Crescendo-angina" OR "Instabiel Angina" OR "Onstabiele Angina" OR "Angina Pectoris in Rust" OR "Angina Pectoris Instabiel" OR "Instabiele Angina Pectoris" OR "Angina Pectoris Onstabiele" OR "Onstabiele Angina Pectoris" OR "Angina Unstable" OR "Crescendo Angina" OR "Worsening Angina" OR "Unstable Angina Pectoris", OR "Crescendo Angina Pectoris" OR "Angina bij Rust" OR "IAP"
Prior PCI	[ANSWER: No] "Geen PCI" OR "PCI niet geindiceerd" OR "Geen indicatie PCI" OR "Geen indicatie voor een PCI" OR "PCI nee" OR "Geen percutane coronaire interventie" [ANSWER: Yes] "Percutane Coronaire Interventie" OR "Percutaneous Coronary Intervention" OR "Percutaneous Coronary Revascularization" OR "pci" OR "bms" OR "bvs" OR "des" OR "stent" OR "dotter"
Prior CABG	("LIMA" OR "RIMA") AND ("CBG" OR "CABG" OR "Coronaire Bypass" OR "Aortocoronaire Bypass" OR "Coronaire Bypassoperatie" OR "CBG" OR "CABG" OR "Coronary Bypass" OR "Coronary Bypass Graft" OR "Coronary Artery Bypass" OR "Coronary Artery Bypass Graft" OR "Coronary Artery Bypass Grafting")
Atrial fibrillation	[ANSWER: No] "Geen AF" OR "AF nee" OR "Geen atrium fibrilleren" OR "atriumfibrilleren nee" OR "geen atriumfibrilleren" OR "atrium fibrilleren nee" OR "geen boezemfibrilleren" OR "boezemfibrilleren nee" [ANSWER: PAF] "PAF" OR "Paroxysmale Atriumfibrillatie" OR "Paroxysmaal Atriumfibrilleren" OR "Atriumfibrilleren Paroxysmaal" OR "PAF" OR "Paroxysmal Af" OR "Paroxysmal Atrial Fibrillation" OR "Paroxysmaal AF" OR "Paroxysmaal atrium fibrilleren" OR "Paroxysmaal boezemfibrilleren" OR "Paroxysmaal boezem fibrilleren" [ANSWER: Yes] "persisterend atriumfibrilleren" OR "persisterend atrium fibrilleren" OR "persisterend AF" OR "persisterend boezemfibrilleren" OR "persisterend boezem fibrilleren" OR "Afib" OR "Atriumfibrillen" OR "Atriumfibrileren" OR "Atriumfbrilleren" OR "Atriumfibrillatie" OR "Atriumfibrilleren" OR "Fibrilleren Atrium" OR "Atrium Fibrilleren" OR "Atriale Fibrillatie" OR "Afib" OR "Atrial Fibrillation" OR "AF"

Table S1b: *(continued)*

Variable	Search terms
Gout	[ANSWER: No] "Geen jicht" OR "Jicht nee" [ANSWER: Yes] "Jicht" OR "Jicht bij" OR "bij Jicht" OR "Arthritis Urica"
Aspirin	DRUG "Aspirin" OR "acetylsalicylzuur" OR "ASA" OR "Acetylsalicylic Acid" OR "Ascal" OR "carbasalaatcalcium"
Oral anticoagulant	DRUG "acenocoumarol" OR "Acenocoumarin" OR "Sintrom" OR "fenprocoumon" OR "marcoumar" OR "Marcumar" OR "apixaban" OR "Eliquis" OR "dabigatran" OR "Pradaxa" OR "Pradax" OR "edoxaban" OR "rivaroxaban"
Statin	DRUG "rosuvastatine" OR "atorvastatine" OR "pravastatine" OR "Eptastatin" OR "simvastatine" OR "fluvastatine" OR "crestor" OR "Lipitor" OR "selektine" OR "Lipostat" OR "Pravachol" OR "Elisor" OR "zocor" OR "Statine"
Ezetimibe	DRUG "ezetimibe" OR "Ezetimib" OR "ezetrol"
ACE inhibitor OR ARB	DRUG "captopril" OR "lisinopril" OR "enalapril" OR "fosinopril" OR "perindopril" OR "ramipril" OR "coversyl" OR "Zestril" OR "candesartan" OR "Irbesartan" OR "Losartan" OR "valsartan" OR "cozaar" OR "aprovel" OR "diovan" OR "pril"
Beta blocker	DRUG "atenolol" OR "bisoprolol" OR "metoprolol" OR "propranolol" OR "selokeen" OR "Bisoprolol" OR "emcor" OR "Atenolol" OR "Sotalol"
Calcium channel blocker	[ANSWER: No (periprocedural verapamil)] DRUG "periprocedural" AND "Verapamil" [ANSWER: Yes] DRUG "diltiazem" OR "amlodipine" OR "nifedipine" OR "isoptin" OR "tildiem" OR "Norvasc" OR "Istin" OR "Amlor" OR "Adalat" OR "Amlodipine"
P2Y12 inhibitor	DRUG "Clopidogrel" OR "Brilique" OR "Effient" OR "Ticagrelor" OR "Plavix" OR "Prasugrel"

Table S2a. Distributions of variables per participating medical center

Variable	Hospital A, % of true	Hospital B, % of true	Hospital C, % of true
Sex (Male)	86.2	84.3	80.7
Current smoker (Yes)	18.9	8.6	11.9
Hypertension (Yes)	85.5	74.3	78.1
Diabetes (Yes)	17.6	31.4	19.3
Insulin dependent diabetes (Yes)	13.8	12.9	13.2
Renal function (No impairment)	93.1	90.0	85.9
Prior ACS (Yes)	92.5	91.4	79.4
Prior PCI (Yes)	93.1	82.9	93.2
CABG (Yes)	9.4	14.3	12.9
AF (Yes)	10.7	10.0	6.1
Gout (Yes)	6.3	5.7	8.7
As or PI (Yes)	99.4	95.7	98.4
Anticoagulant (Yes)	20.1	11.4	26.7
No APT or OAC (Yes)	0.6	1.4	1.3
Statin (Yes)	96.2	78.6	94.9
Ezetimibe (Yes)	20.8	42.9	25.4
ACE inhibitor (Yes)	84.3	74.3	77.8
Beta blocker (Yes)	95.6	92.9	89.7
Calcium channel blocker (Yes)	50.3	42.9	36.7

Table S2b: Accuracy of variables per participating medical center

Variable	Hospital A, accuracy %	Hospital B, accuracy %	Hospital C, accuracy %
Sex	98.5	100.0	100.0
Current smoker	85.8	83.1	85.3
Hypertension	64.2	52.5	63.8
Diabetes	97.0	88.1	95.4
Insulin dependent diabetes	90.3	93.2	90.6
Renal function	95.5	100.0	87.9
Prior ACS	85.8	79.7	84.4
Prior PCI	93.3	88.1	93.5
CABG	98.5	96.6	97.4
AF	88.8	88.1	85.0
Gout	93.3	91.5	92.8
As or PI	80.6	86.4	60.3
Anticoagulant	90.3	94.9	91.9
No APT or OAC	99.3	100.0	98.4
Statin	90.3	76.3	87.9
Ezetimibe	85.8	89.8	87.0
ACE inhibitor	89.6	94.9	86.0
Beta blocker	66.4	84.7	74.6
Calcium channel blocker	75.4	78.0	83.7

Table S2c: Contingency tables of variables per participating medical center

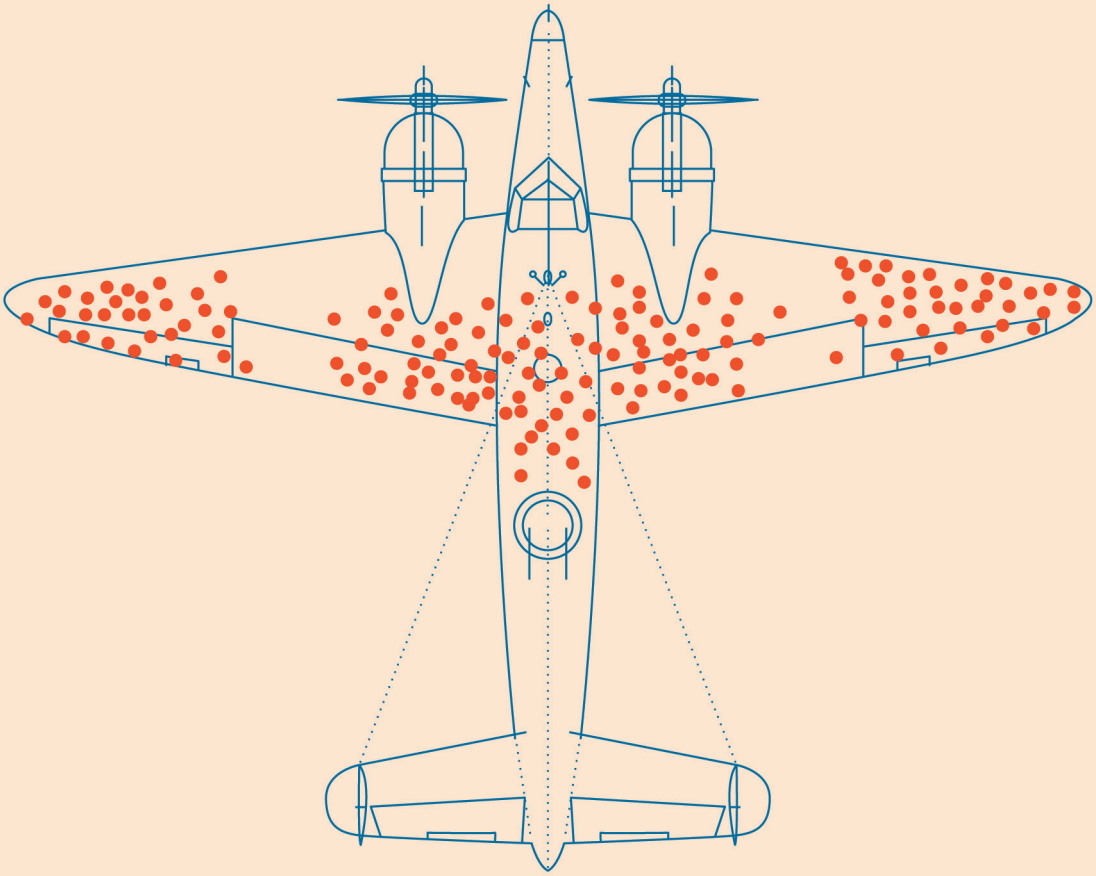
	Hospital A				Hospital B				Hospital C			
Variable	**True in trial, true in EHR (n)**	**False in trial, true in EHR (n)**	**True in trial, false in EHR (n)**	**False in trial, false in EHR (n)**	**True in trial, true in EHR (n)**	**False in trial, true in EHR (n)**	**True in trial, false in EHR (n)**	**False in trial, false in EHR (n)**	**True in trial, true in EHR (n)**	**False in trial, true in EHR (n)**	**True in trial, false in EHR (n)**	**False in trial, false in EHR (n)**
Sex	115	0	2	17	49	0	0	10	248	0	0	59
Current smoker	12	12	7	103	2	4	6	47	17	20	25	245
Hypertension	70	44	4	16	18	25	3	13	152	90	21	44
Diabetes	24	2	2	106	14	6	1	38	46	10	4	247
Insulin dependent DM	8	13	0	113	4	4	0	51	15	26	3	263
Renal function	122	4	2	6	53	0	0	6	253	10	27	17
Prior ACS	110	12	7	5	43	11	1	4	214	31	17	45
Prior PCI	117	7	2	8	46	5	2	6	268	18	2	19
CABG	14	0	2	118	5	2	0	52	34	5	3	265
AF	6	8	7	113	2	5	2	50	13	6	40	248
Gout	6	2	7	119	2	1	4	52	12	15	7	273

Table S2c: Contingency tables of variables per participating med*ical center*

	Hospital A				Hospital B				Hospital C			
Variable	**True in trial, true in EHR (n)**	**False in trial, true in EHR (n)**	**True in trial, false in EHR (n)**	**False in trial, false in EHR (n)**	**True in trial, true in EHR (n)**	**False in trial, true in EHR (n)**	**True in trial, false in EHR (n)**	**False in trial, false in EHR (n)**	**True in trial, true in EHR (n)**	**False in trial, true in EHR (n)**	**True in trial, false in EHR (n)**	**False in trial, false in EHR (n)**
As or PI	108	25	1	0	50	7	1	1	183	119	3	2
Anticoagulant	13	12	1	108	4	3	0	52	56	24	1	226
No APT or OAC	0	1	0	133	0	0	0	59	0	4	1	302
Statin	121	7	6	0	45	4	10	0	267	24	13	3
Ezetimibe	16	11	8	99	22	5	1	31	54	24	16	213
ACE inhibitor	99	13	1	21	41	2	1	15	204	37	6	60
Beta blocker	83	44	1	6	48	9	0	2	203	72	6	26
Calcium channel blocker	37	32	1	64	14	13	0	32	73	39	11	184

Table S2d: Overall test performance scores of variables

Variable	**Sensitivity, %**	**Specificity, %**	**Positive predictive value, %**	**Negative predictive value, %**	**F1-score**
Sex	99.5	100.0	100.0	97.7	0.998
Current smoker	44.9	91.6	46.3	91.2	0.456
Hypertension	89.6	31.5	60.2	72.3	0.720
Diabetes	92.3	95.6	82.4	98.2	0.870
Insulin dependent diabetes	90.0	90.9	38.6	99.3	0.540
Renal function	93.7	67.4	96.8	50.0	0.952
Prior ACS	93.6	50.0	87.2	68.4	0.903
Prior PCI	98.6	52.4	93.5	84.6	0.960
CABG	91.4	98.4	88.3	98.9	0.898
AF	30.0	95.6	52.5	89.3	0.382
Gout	52.6	96.1	52.6	96.1	0.526
As or PI	98.6	1.9	69.3	37.5	0.814
Anticoagulant	97.3	90.8	65.2	99.5	0.781
No APT or OAC	N/A	99.0	N/A	99.8	N/A
Statin	93.7	7.9	92.5	9.4	0.931
Ezetimibe	78.6	89.6	69.7	92.3	0.739
ACE inhibitor	97.7	64.9	86.9	92.3	0.920
Beta blocker	97.9	21.4	72.8	82.9	0.835
Calcium channel blocker	91.2	76.9	59.6	95.9	0.721



Chapter 8

Text-mining electronic healthcare records for trial endpoint data collection

Wouter B van Dijk*, Pieter van der Veere*, Ewoud Schuit, Rolf H H Groenwold, Rieke van der Graaf, Sara Laurijssen, Martine C de Vries, Jeroen Schaap, Erik Badings, Salem H K The, Marco Alings, Diederick E Grobbee, Arend Mosterd, Aernoud T L Fiolet

Draft.

* Authors contributed equally

ABSTRACT

Background

If routinely collected electronic healthcare data are accurate enough to serve as trial endpoint data, leveraging these data may reduce costs of clinical research and facilitate trials on a larger scale . This study aimed to investigate the accuracy of using routinely collected healthcare data from electronic healthcare records (EHRs) to identify major cardiovascular events for in a multicentre randomised clinical trial.

Methods

Participants from a multicentre cardiovascular outcome trial were identified in EHRs of four trial sites. Cardiovascular endpoints were collected using semi-automated text-mining software. The primary endpoint was the composite of cardiovascular death, myocardial infarction, or ischemic stroke (MACE). Secondary endpoints were the composite of MACE plus ischemia-driven coronary revascularization (MACE+) and the individual components of the composite endpoints. Data obtained by EHR data retrieval and investigator-reported endpoint data collection were compared in terms of Kaplan-Meier curves of cumulative event-free survival probabilities.

Results

For the 945 trial participants, 154 endpoints were reported by trial investigators. EHR data retrieval identified 130 (84.4%) of these endpoints: 47 of 51 MACE endpoints (sensitivity 92.2%, specificity 86.5%), and 94 of 114 MACE+ endpoints (sensitivity 82.5%, specificity 81.8%). Cardiovascular death was identified in all six cases (100%), myocardial infarctions in 36 of 38 cases (sensitivity 94.7%, specificity 85.8%), ischemic strokes in five of seven cases (sensitivity 71.4%, specificity 99.3%), and ischemia-driven coronary revascularizations in 47 of the 63 cases (sensitivity 74.6%, specificity 96.7%). Positive predictive values ranged widely and positive predictive values were high. Similar results were found using adjudicated endpoints. Cumulative survival probabilities based on both methods of data collection showed high agreement.

Conclusion

Routinely collected healthcare data from EHRs could be used to accurately collect strictly defined endpoints for a cardiovascular outcome trial. Using routinely collected healthcare data may complement data collection in clinical drug research.

INTRODUCTION

Drug research and development comes at high costs, among others due to the scale of randomized controlled trials (RCTs) with clinical outcomes. Especially in populations with low to moderate absolute risk for study endpoints, over 10,000 participants may be required to detect a clinically relevant treatment effect. Costs of such trials can rise up to 45,000 US dollars per participant.[1,2] The number of new drugs approved relative to the spending on research and development consequently halves every ten years, compromising innovation in drug research.[3,4] Lowering costs in drug research could reduce barriers for all trial phases, facilitate head-to-head comparisons of treatments, and improve patient care.[5]

Data collection is a major cost component in trial conduct, among others driven by personnel costs and site monitoring costs.[6–9] Current practice in trials is to collect data during planned patient-clinician interactions, after which data are registered manually in (electronic) clinical report forms. Most of this information is already collected in electronic healthcare records (EHRs) during routine practice.[9] Extracting information from EHRs may complement or even substitute some parts of data collection in trials.[9,10] However, the use of EHR data in RCTs is limited to date.[11]

A prerequisite to use EHR extracted information in trials is the accurate recognition and mining of this data.[9] Commonly, accuracy of endpoint data collection in clinical trials is safeguarded by strict monitoring, and reporting of potential events by patients and investigators. Best practice prescribes blinded adjudication by an adjudication committee to homogenize endpoint assesment.[12] While studies have investigated the accuracy of insurance and registry data for cardiovascular outcome trials, investigations of the accuracy of data extracted from EHRs are limited.[13,14]

This study aimed to investigate the accuracy of using routinely collected healthcare data from EHRs to identify major cardiovascular events for a multicenter randomized clinical trial.

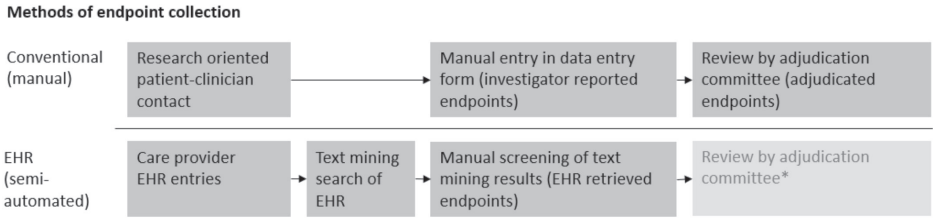
METHODS

This study compared conventional and semi-automated EHR text-mining methods for the collection of endpoint data as a post hoc analysis of the Low-Dose Colchicine 2 (LoDoCo2) trial (**Figure 1**).

Study design and population

The LoDoCo2 trial was a multicenter, randomized, placebo-controlled, clinical trial investigating efficacy of low-dose colchicine in patients with coronary disease. Patients were recruited between August 2014 and December 2018. Methods and main results of the comparison trial were published before.[15,16] For the present study, patients recruited by four sites in the Netherlands were selected. The site selection was based on the total number of included trial participants and different EHR software vendors of the sites. The four selected sites used three different EHR software vendors: Epic Systems Corporation (https://www.epic.com), Chipsoft (htttps://www.chipsoft.com), and CSC Care solutions (https://www.easycare.software). These three EHR software vendors are used in 80% of Dutch hospitals, 40% of North American hospitals and 10% of all hospitals using EHRs globally.[17,18]

Figure 1: Methods of endpoint data collection



*Not performed in this study. Abbreviation: EHR: Electronic health record.

Data collection

For this study, EHR records were searched with a semi-automated EHR text-mining software program to identify the occurrences of any of the pre-defined trial endpoints. Semi-automated refers to the combination automatic detection and manual validation of the results. Endpoints reported in physical investigator-patient interviews during the original trial were used as reference standard.

EHR data retrieval was performed with a commercially available software program (CTcue, version 2.0.12). The program indexes and searches through all structured and free-text EHR data entries. This included clinical notes, reports, diagnosis letters, laboratory measurements, imaging results and histopathology reports. Identification of endpoints was based on a pre-defined query that was developed by three authors (WD, PV and AF) using data from one site. The query consisted of terms and synonyms for all trial endpoints,

and if applicable, blood measurements, drug use, and/or insurance codes. No site-specific optimizations were added to the query, except for the identification of trial participants and identification of drugs administration using local differences. All EHR entries were searched, except for entries that were made for the purpose of the trial.

Trial participants were identified by the software program based on their unique trial identification number as recorded in the EHR. Only EHR entries after trial enrolment were included. All potential endpoints yielded by the text-mining software were manually validated by two blinded investigators (PV and WD).

The endpoints collected for the original trial functioned as reference standard. All potential (i.e., unadjudicated) endpoints during the trial were collected during patient-investigator interviews according to the original trial protocol.[15,16] The participant-investigator interviews took place during clinical evaluations and were scheduled at six-month intervals until the completion of the trial. All follow-up assessments were performed in person, if possible, or by telephone. The interviews were conducted following a standardized electronic case report form, with emphasis on any hospital admission in between visits. All investigator-reported endpoints from the trial were subjected to adjudication by a committee whose members were unaware of the trial-group assignments, following the definitions of an adjudication charter. To reflect actual trial practice, all analyses in the current study were performed using both the investigator-reported endpoints and adjudicated endpoints as reference.

Outcome measures

The primary endpoint was a composite of cardiovascular death, myocardial infarction, or ischemic stroke (major adverse cardiovascular events, or MACE). The secondary endpoint in the current analyses was a composite of cardiovascular death, myocardial infarction, ischemic stroke, or ischemia-driven coronary revascularization (the primary composite endpoint plus ischemia-driven coronary revascularization, or MACE+). Other secondary endpoints were all-cause mortality, the individual components of the composite primary and secondary composite outcome, venous-thrombotic embolism, atrial fibrillation, and diabetes mellitus. Definitions used for the outcomes were similar to those described in the endpoint adjudication charter of the original protocol.[15,16]

Data analysis

Accuracy of endpoint collection by EHR data retrieval was investigated by calculating sensitivity and specificity using investigator-reported endpoints and adjudicated end points as reference standard. Additionally, positive and negative predictive values and Cohen's kappa values were calculated. The main analysis included all participants from the four participating sites.

To quantify the impact of potential misclassification of endpoints when using EHR data, event-free survival times were compared for both endpoint collection methods. Time-to-event was defined as the difference between the date of occurrence of the endpoint and the date of randomization. For patients without an endpoint, follow-up time was censored at the end of trial. Endpoints that occurred before randomization or after the close-out visit were excluded from the analysis. For all endpoints, Kaplan–Meier estimates, and curves of the event-free survival probabilities were calculated to evaluate the timing of endpoint occurrences using the two data collection methods. For atrial fibrillation and diabetes mellitus no time to event data were collected, in concordance with the original trial design. For all tests, the threshold for statistical significance was set at a p-value of 0.05. Multiple testing was not corrected for. Per the design of the current study, all analyses were done irrespective of treatment allocation. All analyses were performed using R software, version 4.0.3.[19]

Ethical oversight, author contributions and conflicts of interest

All patients provided written informed consent to participate in the original trial. The trial protocol was approved by a centralized institutional review board (Medical Research Ethics Committees United, Nieuwegein) and registered in a clinical trial registry (LoDoCo2 Australian New Zealand Clinical Trials Registry number, ACTRN12614000093684). All authors contributed to the design of the study. PV, WD, MA, AM, EB, AM and AF collected and managed the data. The text-mining software was provided free of cost to the investigators. The vendors of the text-mining and EHR software had no role in the design or writing of the protocol or manuscript. All authors had unrestricted access to the data and vouch for the completeness and accuracy of the data and analyses.

RESULTS

Population

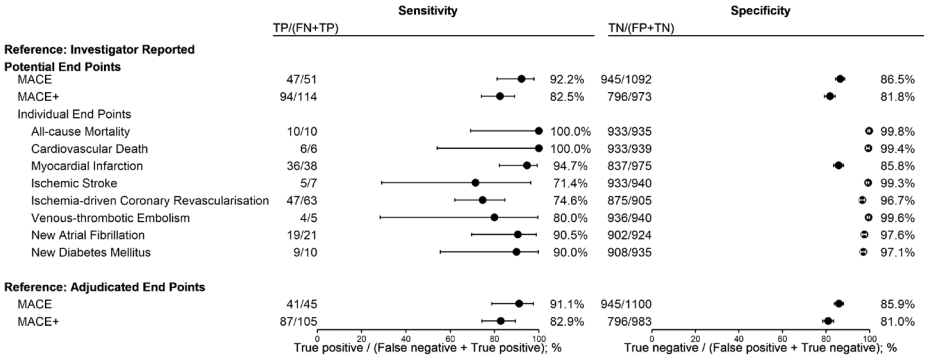
In 939 of the 945 (99.4%) trial participants the trial identification number was registered in the EHR. The mean age of the participants was 65.8 years (Standard Deviation [SD] 8.5 years), 156 (16.5%) were female, 176 (18.6%) had diabetes, and 777 (82.2%) had a history of acute coronary syndrome (**Table S1, Supplementary Materials**).

Accuracy

EHR data retrieval identified 47 of the 51 investigator-reported primary composite endpoints (MACE; sensitivity 92.2%, specificity 86.5%) and 94 of the 114 investigator-reported secondary composite endpoints (MACE+; sensitivity 82.5%, specificity 81.0%; **Figure 2**). All-cause mortality was recorded in ten of ten patients (sensitivity 100%, specificity 99.8%) and cardiovascular death in six of six patients (sensitivity 100%, specificity 99.4%). EHR data retrieval identified 36 of 38 occurrences of myocardial infarction (sensitivity 94.7%, specificity 85.8%), five of seven occurrences of ischemic stroke (sensitivity 71.4%, specificity 99.3%), and 47 of 63 occurrences of ischemia-driven coronary revascularization (sensitivity 74.6%, specificity 96.7%).

EHR data retrieval identified four of five occurrences of venous-thrombotic embolisms (sensitivity 80.0%, specificity 99.6%), 19 of 21 occurrences of atrial fibrillation (sensitivity 90.5%, specificity 97.6%) and nine of ten cases of occurrences of diabetes mellitus (sensitivity 90.0%, specificity 97.1%). Positive predictive values were low to moderate with a wide range between composite and individual endpoints (range 20.7% to 83.3%). Negative predictive values were high with a small range (range 97.5% to 100%; **Table S2, Supplementary Materials**).

Figure 2: Sensitivity and specificity of endpoint identification using EHR data retrieval



The primary endpoint (MACE) was the composite of cardiovascular death, myocardial infarction, or ischemic stroke. The secondary endpoint (MACE+) was the composite of cardiovascular death, myocardial infarction, ischemic stroke, or ischemia-driven coronary revascularization. The formula for sensitivity is: true positive / (false negative + true positive). The formula for specificity is: true negative / (false positive + true negative). The whiskers of the points depict the 95% confidence interval. Abbreviations: FN: false negative; FP: false positive; TN: true negative; TP: true positive; MACE: major adverse cardiovascular event

A total of 114 of the 154 (74.5%) investigator-reported endpoints were confirmed as endpoint after endpoint adjudication by the endpoint adjudication committee. Additional analyses using adjudicated endpoints as reference yielded similar estimates of accuracy (**Table S2, Supplementary Materials**).

Event-free survival times

Cumulative incidence rates for MACE assessed using EHR data retrieval corresponded with cumulative incidence rates based on investigator-reported endpoints (1.9 events per 100 person-years using EHR data retrieval and 2.1 events per 100 person-years using investigator-reported endpoints; **Figure 3**). Similar agreement was observed for MACE+ (2.9 events per 100 person-years using EHR data retrieval and 3.3 events per 100 person-years using investigator-reported endpoints).

Cumulative incidence rates for all-cause mortality did not differ between both methods of collection (0.48 events per 100 person-years). Similar agreement was found for the cumulative incidence rates for myocardial infarction (1.4 events per 100 person-years using EHR data retrieval and 1.5 events per 100 person-years using investigator-reported endpoints) and ischemic stroke (0.24 events per 100 person-years using EHR data retrieval and 0.29 events per 100 person-years using investigator-reported endpoints). Cumulative

incidence rates for ischemia-driven coronary revascularization were estimated lower using EHR retrieval (1.8 events per 100 person-years) as compared to the investigator-reported endpoints (2.3 events per 100 person-years).

Not identified endpoints

A total of 24 investigator-reported endpoints were not identified in the process of EHR data retrieval and manual validation (**Table 1**). Reasons for missing endpoints were absence of EHR entries of the endpoint (6 endpoints), misidentification of the endpoint by EHR data retrieval (2 endpoints) and misclassification of endpoints yielded by EHR data retrieval during manual validation of results (16 endpoints). The distribution of reasons for missed endpoints was similar for the adjudicated endpoints.

Figure 3: Cumulative incidence of the primary endpoint (MACE) and secondary endpoint (MACE+)

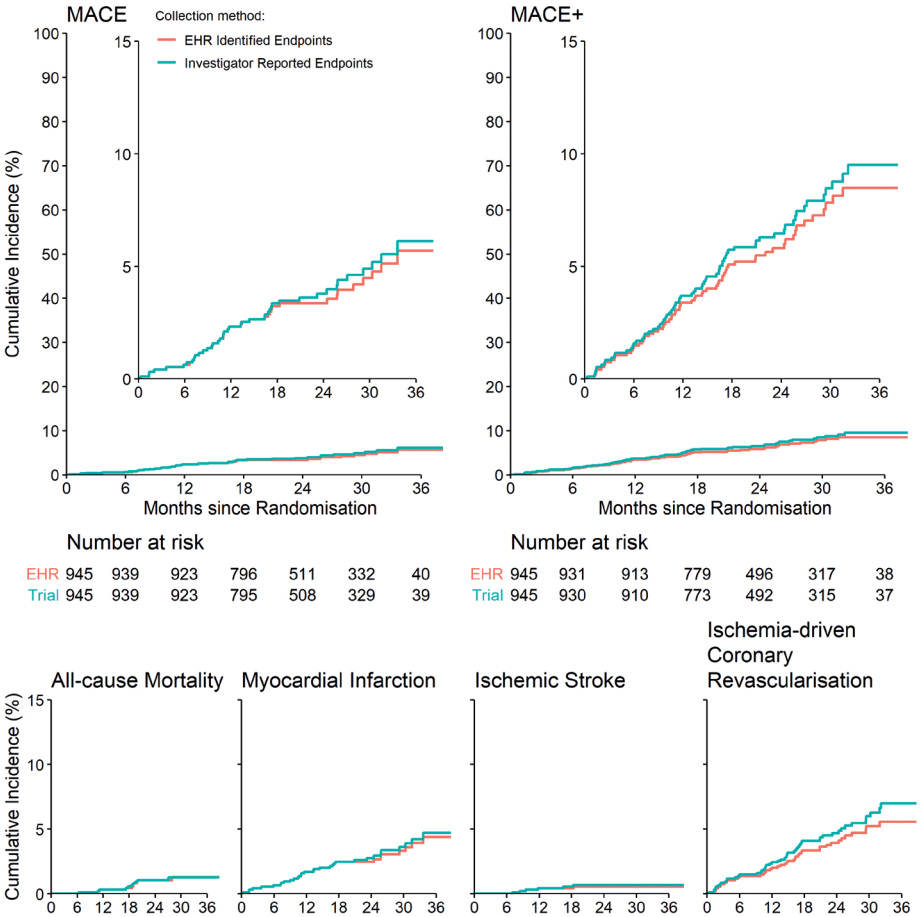


Table 1: Not identified endpoints and adjudicated endpoints

Not identified endpoints	Investigator-reported endpoints (n)	Adjudicated endpoints (n)
No EHR entry for endpoint	6	4
Endpoint not identified by EHR data retrieval (EHR entry for endpoint was present)	2	0
Endpoint identified by EHR data retrieval, but missed during manual EHR validation	16	14

Abbreviations: EHR = Electronic Healthcare Records

DISCUSSION

This analysis of the data of a randomized clinical outcome trial investigated the accuracy of endpoint identification using routinely collected EHR data as compared to conventional investigator-reported endpoints by patient interviews. The most notable findings include the following:

First, we found that almost all trial participants could be identified using an EHR search algorithm. Second, we found high sensitivity and negative predictive values for this method to detect the composite of the clinically relevant composite endpoint of cardiovascular death, myocardial infarction, and ischemic stroke (MACE), and the individual components of this endpoint. However, the method had a lower sensitivity to detect the more ambiguously registered endpoint of ischemia-driven revascularization. Positive predictive values were low to moderate for all endpoints. Third, we found strong agreement in estimated cumulative incidence rates between the two methods of collection. Using investigator-reported endpoints or adjudicated endpoints as reference standard did not affect EHR data retrieval accuracy. Fourth, we found that most misidentified endpoints depended on misclassification by the manual operator.

Our method of validating routinely collected healthcare data from hospital EHRs from multiple vendors, using data from a clinical outcomes trial as reference standard, has not been described before. The results are similar to previous studies that compared correspondence of routinely collected healthcare data registered in healthcare claim databases to investigator-reported outcomes and endpoint evaluation by an endpoint adjudication committee, in which sensitivity of using routinely collected healthcare data ranged between 49% to 100% [13,20–23]

Our results show sensitivity, specificity, and negative predictive values are particularly high for strictly defined clinical outcomes, such as major adverse cardiovascular events. With these properties, EHR data retrieval from routinely collected healthcare data can be considered to complement or substitute investigator-reported endpoint collection of such outcomes in clinical research. Less strictly defined or ambiguously registered clinical outcomes such as ischemia-driven revascularization show lower levels of accuracy. EHR data retrieval using broadly defined clinical outcomes is appropriate for clinical research involving large datasets, where estimates are less effected by missed observations.[24] Depending on the context of the clinical research, methods that increase sensitivity could reduce specificity, as was observed in our study with the secondary composite endpoint and myocardial infarction, and low positive predictive values for all endpoints. For the use of clinical endpoints, an adjudication committee could help to increase specificity in such circumstances.

Endpoint identification using EHR text-mining comes with three consequences for scientific research, and clinical trials in particular.

First, using EHR-based endpoint information instead of investigator-reported endpoints will usually lower the power of a clinical trial, provided misclassification of endpoints is independent of treatment arms. As sensitivity and specificity of the EHR-based endpoints depart from the ideal values of 1, the power of the trial will become lower. To maintain a pre-specified power, this would typically mean that sample sizes need to increase when using EHR-based endpoints instead of using investigator-reported endpoints. In particular low-cost, pragmatic investigations, such as in low-income countries or head-to-head comparisons for on-market treatments might benefit from EHR data collection methods. These studies meet clinical needs, but are often difficult to fund.[25,26]

Second, missed endpoints may introduce bias in assessing treatments. For example, treatments with more adverse effects than control might lead to higher number of hospital visits and increased chances to register endpoints than in the control arm, spuriously increasing risk in the treatment arm. However, such bias also applies to a smaller extent for regular endpoint collection as trial endpoints are also identified in routine physician-patient interactions.

Third, EHR data collection is limited to information that is regularly collected. Data that is not regularly ascertained in routine care, such as rare adverse events, or have ambiguous registration will be more difficult to identify in EHR data retrieval accurately.

Study limitations

We acknowledge several limitations. First, the results in this study were yielded using a closed-source, commercial software package. As a result, we could not validate underlying algorithms used in the software packages employed directly. To mitigate this limitation, we assessed multiple EHR vendors. Second, the medical background and training of the operator in validating the results will affect sensitivity. We could not isolate the magnitude of this effect on our estimates. Lastly, the limited sample size in this analysis and the low incidence rate of some endpoints in the trial affected precision of estimates sensitivity.

CONCLUSIONS

EHR data retrieval from routinely collected healthcare data could be used to accurately collect strictly defined endpoints for a cardiovascular outcome trial when compared to regular investigator-reported endpoints using patient interviews. Both methods yielded similar cumulative incidence rates and findings are irrespective of endpoint adjudication. Positive predictive values ranged widely, and less strictly defined endpoints resulted in lower levels of sensitivity. Using routinely collected healthcare data should be considered in data collection of clinical drug research.

REFERENCES

1. Speich B, von Niederhäusern B, Schur N, et al. Systematic review on costs and resource use of randomized clinical trials shows a lack of transparent and comprehensive data. J Clin Epidemiol. 2018;96:1-11.
2. Moore TJ, Zhang H, Anderson G, Alexander GC. Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015-2016. JAMA Intern Med. 2018;178(11):1451.
3. Van Norman GA. Overcoming the Declining Trends in Innovation and Investment in Cardiovascular Therapeutics: Beyond EROOM's Law. JACC Basic to Transl Sci. 2017;2(5):613-625.
4. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. 2012;11(3):191-200.
5. Djurisic S, Rath A, Gaber S, et al. Barriers to the conduct of randomised clinical trials within all disease areas. Trials. 2017;18(1):360.
6. Sertkaya A, Birkenbach A, Berlind A, Eyraud J. Examination of Clinical Trial Costs and Barriers for Drug Development. Lexington; 2014.
7. Munos B. Lessons from 60 years of pharmaceutical innovation. Nat Rev Drug Discov. 2009;8(12):959-968.
8. Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? Nat Rev Drug Discov. 2017;16(6):381-382.
9. Mc Cord KA, Al-Shahi Salman R, Treweek S, et al. Routinely collected data for randomized trials: Promises, barriers, and implications. Trials. 2018;19(1).
10. van Dijk WB, Fiolet ATL, Schuit E, et al. Text-mining in electronic healthcare records can be used as efficient tool for screening and data-collection in cardiovascular trials: a multicenter validation study. J Clin Epidemiol. 2020;132(1):97-105.
11. Lensen S, Macnair A, Love SB, et al. Access to routinely collected health data for clinical trials - Review of successful data requests to UK registries. Trials. 2020;21(1):398.
12. Vantongelen K, Rotmensz N, Van Der Schueren E. Quality control of validity of data collected in clinical trials. Eur J Cancer Clin Oncol. 1989;25(8):1241-1247.
13. Kjøller E, Hilden J, Winkel P, et al. Agreement between public register and adjudication committee outcome in a cardiovascular randomized clinical trial. Am Heart J. 2014;168(2):197--204.e4.
14. St. Peter WL, Liu J, Weinhandl E, Fan Q. A Comparison of Sevelamer and Calcium-Based Phosphate Binders on Mortality, Hospitalization, and Morbidity in Hemodialysis: A Secondary Analysis of the Dialysis Clinical Outcomes Revisited (DCOR) Randomized Trial Using Claims Data. Am J Kidney Dis. 2008;51(3):445-454.
15. Nidorf SM, Fiolet ATL, Mosterd A, et al. Colchicine in Patients with Chronic Coronary Disease. N Engl J Med. 2020:1-10.
16. Nidorf SM, Fiolet ATL, Eikelboom JW, et al. The effect of low-dose colchicine in patients with stable coronary artery disease: The LoDoCo2 trial rationale, design, and baseline characteristics. Am Heart J. 2019;218:46-56.
17. KPMG. The MedTech Market in the Netherlands.; 2017.
18. Zorgvisie. Het complete epd-overzicht: welk ziekenhuis heeft welke leverancier? Houten: Bohn Stafleu van Loghum 2019.
19. R Core Team. R: A Language and Environment for Statistical Computing. 2020. https://www.r-project.org/.
20. Hlatky MA, Ray RM, Burwen DR, et al. Use of medicare data to identify coronary heart disease outcomes in the women's health initiative. Circ Cardiovasc Qual Outcomes. 2014;7(1):157-162.
21. Barry SJE, Dinnett E, Kean S, Gaw A, Ford I. Are Routinely Collected NHS Administrative Records Suitable for Endpoint Identification in Clinical Trials? Evidence from the West of Scotland Coronary Prevention Study. Pajewski NM, ed. PLoS One. 2013;8(9):e75379.

22. St. Peter WL, Liu J, Weinhandl ED, Fan Q. Linking Centers for Medicare & Medicaid Services data with prospective DCOR trial data: Methods and data comparison results. Hemodial Int. 2008;12(4):480-491.
23. Rodrigues C, Odutayo A, Patel S, et al. Comparison between cardiovascular trial outcomes and treatment effects using clinical endpoint committee adjudication versus routine health data: A systematic review. J Am Coll Cardiol. 2019;73(9):1753.
24. Höfler M. The effect of misclassification on the estimation of association: a review. Int J Methods Psychiatr Res. 2005;14(2):92-101.
25. Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ. 2005;331(7521):897-900.
26. VanLare JM, Conway PH, Sox HC. Five Next Steps for a New National Program for Comparative-Effectiveness Research. N Engl J Med. 2010;362(11):970-973.
27. Kellum JA, Lameire N, Aspelin P, et al. Kidney disease: improving global outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury. Kidney Int Suppl. 2012;2(1):1-138.

SUPPLEMENTAL MATERIALS

Table S1: Baseline characteristics of the study sample compared to the original LoDoCo2 cohort

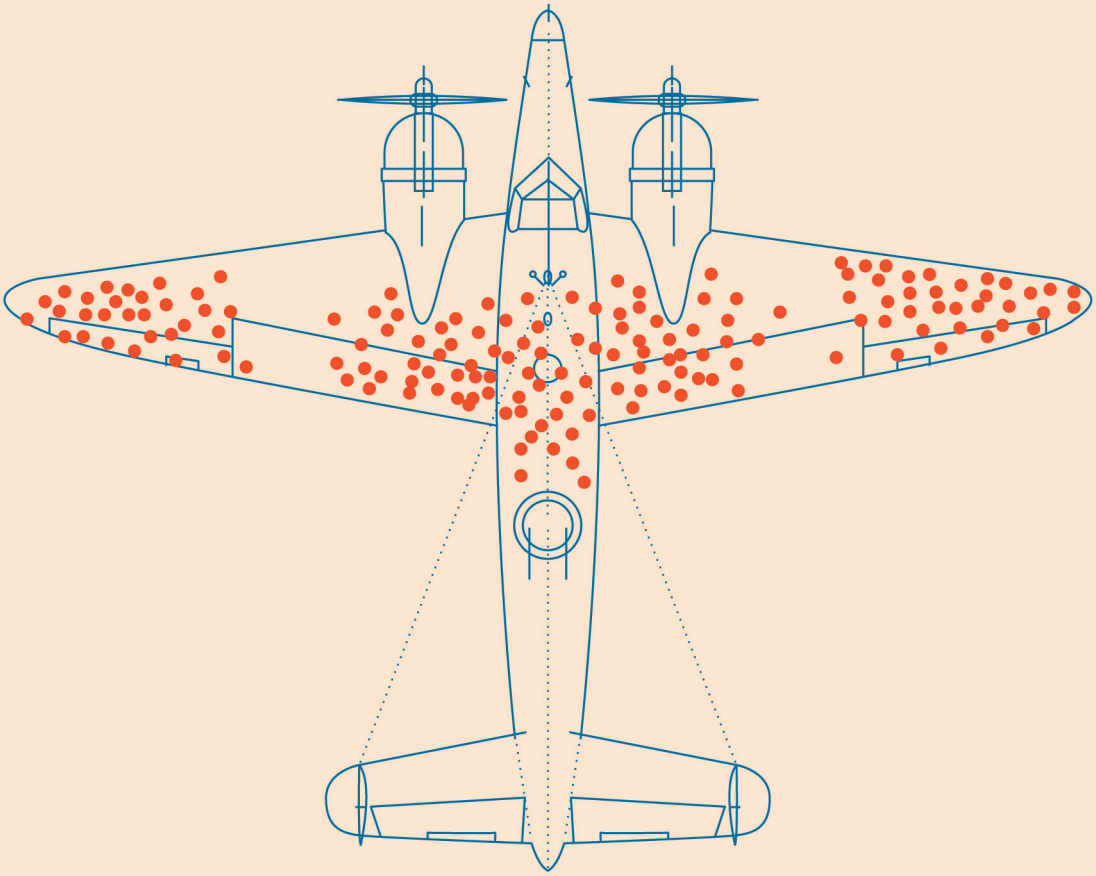
Characteristics	Current study (n = 945)	LoDoCo2 (n = 5522)
Age - yr (SD)	65 (8.5)	66 (8.6)
Female sex - n (%)	156 (16.5)	746 (13.5)
Current smoker - n (%)	148 (15.7)	648 (11.7)
Hypertension - n (%)	482 (51.0)	2,808 (50.9)
Diabetes - n (%)	176 (18.6)	1,007 (18.2)
Kidney function stage 3a* - n (%)	96 (10.2)	306 (5.5)
Prior acute coronary syndrome - n (%)	777 (82.2)	4,658 (84.4)
Prior coronary revascularization	863 (91.3)	4,621 (83.7)
Coronary-artery bypass grafting- n (%)	122 (12.9)	712 (12.9)
Percutaneous coronary intervention- n (%)	798 (84.4)	4,177 (75.6)
History of atrial fibrillation - n (%)	146 (15.4)	649 (11.8)
History of gout - n (%)	86 (9.1)	446 (8.1)
Medication use		
Single antiplatelet therapy - n (%)	645 (68.3)	3,699 (67.0)
Dual antiplatelet therapy - n (%)	171 (18.1)	1,281 (23.2)
Anticoagulant - n (%)	160 (16.9)	673 (12.2)
No antiplatelet agent or anticoagulant - n (%)	1 (0.1)	17 (0.3)
Statin - n (%)	850 (89.9)	5,190 (94.0)
Ezetimibe - n (%)	191 (20.2)	1,071 (19.4)
Any lipid-lowering agent - n (%)	890 (94.2)	5,334 (96.6)
Renin–angiotensin inhibitor - n (%)	672 (71.1)	3,960 (71.7)
Beta-blocker - n (%)	658 (69.6)	3,429 (62.1)
Calcium-channel blocker - n (%)	205 (21.7)	1,244 (22.5)

*Stage 3a kidney function refers to an estimated glomerular filtration rate of 45 to 59 ml per minute per 1.73 m2 (mildly to moderately decreased). Stages are based on the Kidney Disease: Improving Global Outcomes (KDIGO) Clinical Practice Guideline for Acute Kidney Injury.[27]
Abbreviations: SD: standard deviation

Table S2: Accuracy of EHR data retrieval as compared to investigator-reported endpoints and adjudicated endpoints

	TP	FP	TN	FN	Sensitivity (95% CI)	Specificity (95% CI)	PPV	NPV	κ
Reference: Investigator-reported Endpoints									
MACE	47	147	945	4	92.2% (81.1% to 97.8%)	86.5% (84.4% to 88.5%)	24.2%	99.6%	0.34
MACE+	94	177	796	20	82.5% (74.2% to 88.9%)	81.8% (79.2% to 84.2%)	34.7%	97.5%	0.40
Individual Endpoints									
All-cause Mortality	10	2	933	0	100.0% (69.2% to 100.0%)	99.8% (99.2% to 100.0%)	83.3%	100.0%	0.91
Cardiovascular Death	6	6	933	0	100.0% (54.1% to 100.0%)	99.4% (98.6% to 99.8%)	50.0%	100.0%	0.66
Myocardial Infarction	36	138	837	2	94.7% (82.3% to 99.4%)	85.8% (83.5% to 88.0%)	20.7%	99.8%	0.30
Ischemic Stroke	5	7	933	2	71.4% (29.0% to 96.3%)	99.3% (98.5% to 99.7%)	41.7%	99.8%	0.52
Venous-thrombotic Embolism	4	4	936	1	80.0% (28.4% to 99.5%)	99.6% (98.9% to 99.9%)	50.0%	99.9%	0.61
New Atrial Fibrillation	19	22	902	2	90.5% (69.6% to 98.8%)	97.6% (96.4% to 98.5%)	46.3%	99.8%	0.60
New Diabetes Mellitus	9	27	908	1	90.0% (55.5% to 99.7%)	97.1% (95.8% to 98.1%)	25.0%	99.9%	0.38
Ischemia-driven Coronary Revascularization	47	30	875	16	74.6% (62.1% to 84.7%)	96.7% (95.3% to 97.8%)	61.0%	98.2%	0.65
Reference: Adjudicated Endpoints									
MACE	41	155	945	4	91.1% (78.8% to 97.5%)	85.9% (83.7% to 87.9%)	20.90%	99.60%	0.30
MACE+	87	187	796	18	82.9% (74.3% to 89.5%)	81.0% (78.4% to 83.4%)	31.80%	97.80%	0.37

The primary endpoint (MACE) was the composite of cardiovascular death, myocardial infarction, or ischemic stroke. The secondary endpoint (MACE+) was the composite of cardiovascular death, myocardial infarction, ischemic stroke, or ischemia-driven coronary revascularization. Abbreviations: FN: false negative; FP: false positive; MACE: major adverse cardiovascular event; NP; negative predictive value; PPV, positive predictive value; TN: true negative; TP: true positive.



Chapter 9

Dynamics in cardiac surgery: trends in populationcharacteristics and the performance of the EuroSCORE II over time

Wouter B van Dijk, Artuur M Leeuwenberg, Diederick E Grobbee, Sabrina Siregar, Saskia Houterman, Edgar J Daeter, Martine C de Vries, Rolf H H Groenwold, Ewoud Schuit

Submitted.

ABSTRACT

Background

The EuroSCORE II facilitates the assessment of mortality risks in patients undergoing cardiac surgery. Its predecessor, the EuroSCORE showed a notorious decrease in performance over time. The aim of this paper was to investigate whether the EuroSCORE II showed a similar trend over time and to investigate dynamics in predictors included in the model.

Methods

A cohort study was performed using data from the Netherlands Heart Registration (NHR). All cardiothoracic surgical procedures performed between January 1, 2013 and December 31, 2019 were included for analysis. Performance of the EuroSCORE II was assessed on across three-month time intervals in terms of calibration and discrimination. For subgroups of major surgical procedures, performance of the EuroSCORE II was assessed across twelve-month time intervals. Changes in individual EuroSCORE II predictors over time were assessed graphically.

Results

A total of 103,404 cardiothoracic surgical procedures was included in this study. Observed mortality risk ranged between 1.9% and 3.9% across three-month time intervals. The mean predicted mortality risk ranged between 3.7% and 4.5%. The corresponding observed:expected ratios ranged from 0.50 to 0.99. Discriminative performance in terms of the c-statistic ranged between 0.82 and 0.90. The EuroSCORE II persistently overestimated mortality compared to observed mortality, this finding was consistent across all major cardiothoracic surgical procedures.

Distributions of individual predictors varied broadly across predictors over time. Most notable trends were a decrease in elective surgery from 75% to 54% and a rise in patients with no or NYHA I class heart failure from 23% to 33%.

Conclusions

The EuroSCORE II consistently overestimates mortality risks of all types of major cardiothoracic surgical procedures in the Netherlands.

INTRODUCTION

The EuroSCORE II model aims to support clinicians and their patients to determine whether benefits of cardiac surgery outweigh mortality risks associated with these procedures.[1] After implementation in clinical practice prediction models, like the EuroSCORE II, are generally used for many years without any re-assessment of their performance. Consequently, prediction models are at risk of showing decreased performance over time, a phenomenon also known as concept drift.[2] Mortality predictions of the original EuroSCORE model, for instance, were found to increasingly overestimate the mortality risk over time.[3–5] Underlying mechanisms were found in changing patient risk profiles, improved outcomes and the introduction of novel interventions.[5] This drift of the original EuroSCORE was also the main reason to introduce the EuroSCORE II, as in use today.[1]

As prediction models become more frequently in use in clinical practice it is increasingly important to regularly check their performance and update them when needed. The increased interest in registries, real world data, and the concept of learning healthcare systems will likely facilitate more frequent or even continuous assessment of concept drift.[6] As the EuroSCORE II almost celebrates its ten-year anniversary it is time to assess its performance and dynamics in characteristics of the patients for whom the model is potentially applied.

As prediction models become more frequently in use in clinical practice it is increasingly important to regularly check their performance and update them when needed. The increased interest in registries, real world data, and the concept of learning healthcare systems will likely facilitate more frequent or even continuous assessment of concept drift.[6] As the EuroSCORE II almost celebrates its ten-year anniversary it is time to assess its performance and dynamics in characteristics of the patients for whom the model is potentially applied.

Therefore, the aim of our study was to investigate the performance of the EuroSCORE II and dynamics in underlying predictors over time.

METHODS

This study was a cohort study assessing performance trends and population dynamics of the EuroSCORE II over time, using data from the Netherlands Heart Registration (NHR).

STUDY POPULATION

Data for all cardiothoracic surgery procedures performed in adult patients in the Netherlands between January 1st, 2013, and December 31st, 2019, were extracted from the NHR. No distinction was made between first interventions and reinterventions as reinterventions are included as a predictor in the EuroSCORE II.

The NHR comprises prospectively collected data on almost 1.5 million cases of adults with cardiovascular disease and more than 200,000 case of patients who underwent cardiac surgery in one of the sixteen Dutch cardiothoracic medical centers since 2007.[7] Data of the NHR comprise a wide range of patient characteristics (including the original EuroSCORE and EuroSCORE II variables), procedural variables and outcome measures. Data received from participating medical centers collected in accordance with the NHR manual (https://www.nederlandsehartregistratie.nl) reviewed and audited.

Outcome measures

The primary outcome for this study was the performance of the EuroSCORE II model in predicting in-hospital mortality over time.[1] In addition, this study aimed to investigate dynamics of the predictors of the EuroSCORE II model.

EuroSCORE II

In short, the EuroSCORE II model was developed to predict in-hospital mortality of patients after cardiac surgery.[1] The dataset substantiating model development comprised 22,381 cardiothoracic patients form 154 hospitals in 43 countries. Tenfold cross-validation was conducted during model development, dividing the original dataset into ten equally sized random samples of which 90% was used for fitting the model and 10% for validation. Performance of the EuroSCORE II comprised an observed:expected-(O:E)-ratio of 1.06 (observed: 4.18%; expected: 3.95%), and a c-statistic of 0.81 in the validation data. For more details on the EuroSCORE II model we refer to the original publication.[1]

Data analysis

The main analysis included all patients consecutively included in the NHR within the previously defined timeframe. Missing data on individual predictors and the outcome was assessed by calculating overall proportions of missing data, the proportion of missing data across the three-month intervals, and the overall proportion of missing data per individual predictor. A table was constructed to compare patients with information on all

predictors and the outcome available (i.e., without missing values) to patients with one or more missing values on any of the predictors or the outcome. To test for differences between these groups, chi-square tests and t-tests were performed for categorical and continuous predictors, respectively. Missing data were imputed using multiple imputation techniques when required. The imputation model included all predictor variables and the outcome. Data were imputed 10 times to retain potential heterogeneity in model performance or population characteristics over time. Confidence intervals (95%-CIs) were pooled from 200 bootstrap samples evenly distributed over all imputation sets, following Schomaker et al. (2018).[8] All analyses described below were performed separately within each imputation set and pooled afterwards using Rubin's rules.[9]

Model performance was assessed in terms of calibration and discrimination across the three-month intervals. Calibration refers to agreement between the predicted mortality risk and the observed proportions of cardiac surgery patients who died in-hospital. Calibration was assessed graphically in calibration plots and numerically in terms of the O:E-ratio, the calibration-in-the-large (CITL) and the calibration slope. The O:E ratio was calculated and plotted by dividing the incidence of mortality by the average mortality risk as calculated by the EuroSCORE II in the same group of patients. An O:E ratio of one indicates good overall agreement between observed and expected in-hospital mortality. The calibration-in-the-large and calibration slope were derived from a calibration plot. The calibration line in the plot was described with the calibration slope (b) and with an intercept (a), given that the calibration slope is set to 1 (a|b = 1, calibration-in-the-large), as proposed by Cox.[10] With perfect calibration, the calibration-in-the-large would be zero, and the calibration slope would equal one. Discrimination refers to the ability of the model to distinguish between patients who died and patients who survived and was assessed with the area under the receiver-operating characteristic curve (AUC).[11] Trends over time for all model performance measures were examined in plots drawn with non-parametric locally weighted smoothing (LOESS).

Subsequent analogue analyses were performed for subgroups stratified for four major surgical procedures (isolated CABG, aortic valve surgery, mitral valve surgery, aorta surgery) across twelve-month intervals. For these latter analyses, twelve-month intervals were chosen to increase the number of events per sampling interval.

Similar plots were constructed to assess dynamics in individual model predictors. To assess effects of individual predictors on the overall model performance, a plot comprising all individual predictors centered around their initial three months means was produced.

All analyses were performed using R software, version 4.0.3.18.[12]

Ethical aspects

An exemption from the Dutch law on medical research on human beings was obtained from the Medical Ethical Review Committee Utrecht prior to this study (protocol number 20-698/C).

RESULTS

Population

In total, 103,404 procedures of adult patients who received cardiothoracic surgery were recorded in the NHR between January 1, 2013 and December 31, 2019. Overall, 7.90% of data were missing, with the highest percentage of missing values for NYHA class (30.2 %), poor mobility (20.2 %), and Canadian Cardiovascular Society IV angina pectoris score (18.3%) (see **Supplemental Materials, Table S1 and Figure S1** for missing data trends per predictor over time). For 36,947 patients (35.7%) at least one predictor or outcome value was missing. Notably more data were missing in the first two years of the dataset compared to later years. After imputation of missing data, all 103,404 cases were included for further analysis. Three-month segments comprised an average of 4,309 patients.

Overall, patients were 65.8 years of age at the time of surgery, with 27.5% being female, and an average mortality of 2.8% (**Table 1**).

Overall EuroSCORE II performance

Across three-month time intervals, observed mortality of cardiothoracic surgery ranged between 1.9% (95%-CI 1.3-2.4) and 3.9% (95%-CI 3.6-4.2) over time (**Figure 1)**. Expected mortality, calculated by the EuroSCORE II, ranged between 3.7% (95%-CI 3.4-3.9) and 4.5% (95%-CI 4.4-4.5) over time. Plots showed observed and expected mortality to be coinciding closely over time. O:E-ratios ranged from 0.50 (95%-CI 0.45-0.56) to 0.99 (95%-CI 0.92-1.03).

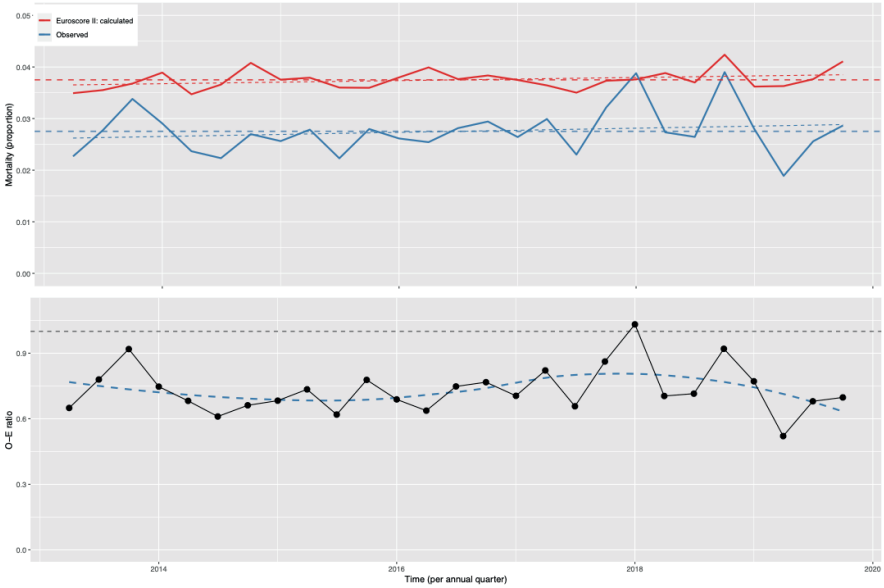
Calibration-in-the-large for the EuroSCORE II were ranging from -0.81 (95%-CI -1.03--0.60) to -0.02 (95%-CI -0.12-0.08) **(Supplemental materials, Figure S2)**. Slopes for the EuroSCORE II centered around one, ranging from 0.10 (95%-CI -0.05-0.25) to 1.21 (95%-CI 1.05-1.38).

Table 1: Overall baseline characteristics of EuroSCORE II predictors (N=103,404)

Characteristic	n (%) or mean (SD)
Outcome	
In-hospital mortality, n (%)	2,678 (2.8)
Predictor	
Age, years (SD)	65.8 (11.0)
Female, n (%)	27,988 (27.5)
Creatinine concentration, mmol/L (SD)	93.2 (47.8)
Dialysis, n (%)	412 (0.40)
Insulin-dependent diabetes mellitus, n (%)	8,169 (7.9)
NYHA, n (%)	
Class I	34,139 (33.0)
Class II	36,563 (35.4)
Class III	26,419 (25.6)
Class IV	6,193 (6.0)
LV-function, % (SD)	51.4 (9.49)
CCS4, n (%)	6,204 (6.0)
Active endocarditis, n (%)	2,698 (2.6)
Extracardiac arteriopathy, n (%)	11,788 (11.3)
Chronic pulmonary disease, n (%)	10,960 (10,6)
Pulmonary artery systolic pressure, mmHg, (SD)	26.6 (6.57)
Recent myocardial infarction, n (%)	22,335 (21.6)
N/M mobility, n (%)	2,554 (2.5)
Urgency, n (%)	
Elective	57,033 (55.2)
Urgent	38,673 (37.4)
Emergency	6,659 (6.4)
Salvage	1,039 (1.0)
Critical preoperative state, n (%)	4,022 (3.9)
Weight of procedure, n (%)	
Isolated CABG	53,432 (51.7%)
1 procedure (non-CABG)	24,827 (24.4)
2 procedures	19,202 (18.6%)
3 or more procedures	5,943 (5.7%)
Thoracic aorta surgery, n (%)	10,065 (9.7)
Previous cardiac surgery (redo), n (%)	7,052 (6.8)

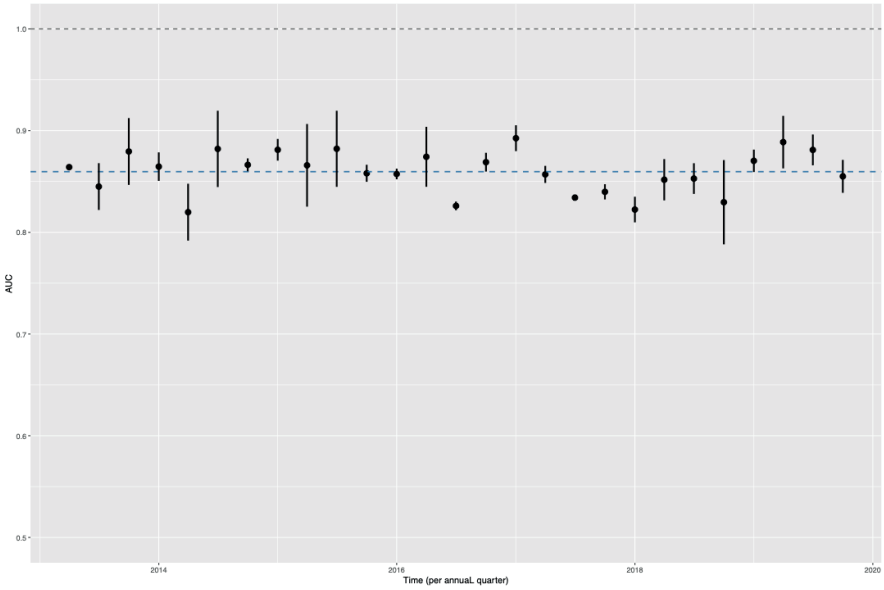
CCS4: CCS class 4 angina; LV function: left ventricular function; N/M mobility: neurological or musculoskeletal dysfunction severely affecting mobility; NYHA: NYHA: New York Heart Association heart failure category; CABG: coronary artery bypass grafting.

Figure 1: Observed and expected mortality of the EuroSCORE II over time



AUCs, of the EuroSCORE II ranged between 0.82 (95%-CI 0.79-0.85) and 0.90 (95%-CI 0.89-0.91). (**Figure 2**). The mean AUC over time was 0.84 (95%-CI 0.81-0.87).

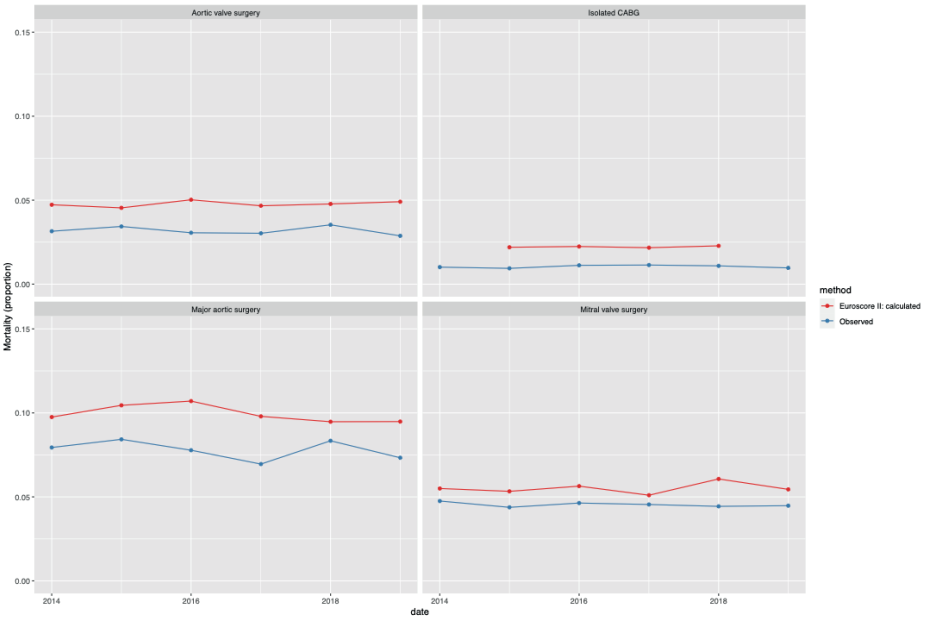
Figure 2: Area under the receiver-operating characteristic curve with 95%-confidence interval of the EuroSCORE II over time



EuroSCORE II performance per major procedure type

For isolated CABG observed mortality ranged from 0.9% (95%-CI 0.7-1.1) to 1.1% (95%-CI 0.9-1.4), and expected mortality from 2.2% (95%-CI 2.2-2.3) to 2.3% (95%-CI 2.3-2.3) (**Figure 3**). For aortic valve surgery observed mortality ranged from 2.9% (95%-CI 2.5-3.2) to 3.5% (95%-CI 2.6-4.5), and expected mortality from 4.8% (95%-CI 4.7-4.8) to 5.3% (95%-CI 5.2-5.4). For mitral valve surgery observed mortality ranged from 4.4% (95%-CI 3.6-5.1) to 4.8% (95%-CI 4.4-5.1), and expected mortality from 5.3% (95%-CI 5.0-5.6) to 6.3% (95%-CI 6.3-6.4). For major aortic surgery observed mortality ranged from 7.0% (95%-CI 5.9-8.4) to 8.4% (95%-CI 6.5-10.5), and expected mortality from 10.1% (95%-CI 10.0-10.2) to 11.4% (95%-CI 10.8-12.0).

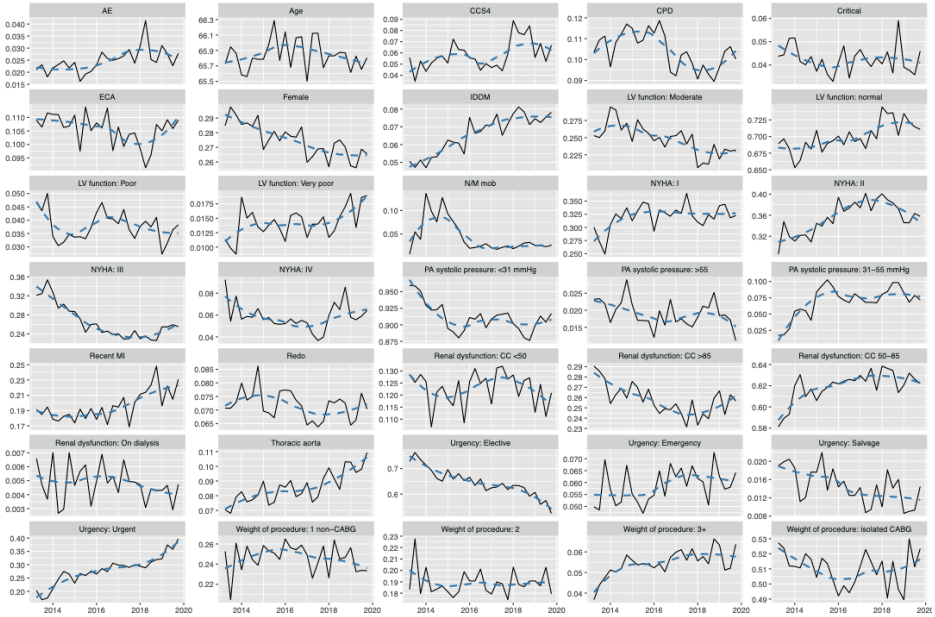
Figure 3: EuroSCORE II. performance per major procedure type



Dynamics of predictors

Trends in individual model predictors varied broadly between predictors (**Figure 4**). Most notable trends were the decrease in elective surgery from 75% to 54%, the rise in patients with no or NYHA I class heart failure from 23% to 33%, the rise in patients with a recent myocardial infarction from 19% to 23%, and the rise in aorta surgery from 7% to 11%.

Figure 4: Trends in individual EuroSCORE II model predictors over time



DISCUSSION

This study examined the model performance of the EuroSCORE II over time since its introduction. Our results show that the EuroSCORE II persistently overestimates in-hospital mortality after cardiothoracic surgery in the Netherlands. Dynamics of individual model predictors showed an increase in urgency of procedures with subsequent decrease in elective procedures, and a decrease in heart failure severity. These changes did not result in a substantial change of the EuroSCORE II's model performance over time.

Since its introduction the performance of the EuroSCORE II has been assessed multiple times for specific populations or in comparison to other risk scores. In general, authors found mixed results for calibration, and satisfactory results for discrimination (i.e. an AUC of >0.7) of the EuroSCORE II.[13–16] For instance, in patients undergoing aortic valve surgery the model was found to perform satisfactory,[16] while in patients undergoing minimal-invasive cardiothoracic surgery calibration was found to be accurate for low risk patients only.[14] In a recent study on risks of cardiothoracic surgery for octogenarians, the EuroSCORE II was found to steadily underestimate mortality risks.[15] When compared to other risk prediction models like the original EuroSCORE, the logistic version of the original

EuroSCORE and Society of Thoracic Surgery (STS)-short term risk score, the EuroSCORE II is typically found to perform better than the first two models and equal to the latter.[13,17] Notably, a single center study validating the EuroSCORE II in a Dutch medical center found the model to systematically underestimate in-hospital mortality, while our study found overestimation of mortality for all major cardiothoracic procedures.[18] A possible explanation for these different research findings is that the single center study was a subset of the data used in this study.

Noticeable also was the high number of missing values for certain predictors like NYHA, and poor mobility. In studies on the performance of the EuroSCORE I with NHR data in the years before the introduction of the EuroSCORE II substantially less predictor values were missing.[4,19] The EuroSCORE II seems to encompass predictors that are less frequently reported by clinicians or that are not applicable for many patients, even though they are relevant for the model. For example, the default value of the NYHA predictor assumes patients to be scored as NYHA I at minimum. However, if patients are not dyspneic it is not possible for clinicians to assign patients a NYHA class at all, which means no NYHA class is reported for that patient.

Future perspectives

Dynamics in populations underlying clinical prediction models are increasingly recognized and acknowledged in research.[22] Still, existing prediction models are only updated infrequently. As previously mentioned, the original EuroSCORE was used for a decade before receiving its much-needed update to the EuroSCORE II. In comparison, introduced in 2008 and updated in 2018, the STS-short term risk score was updated after a decade of use also.[23] Similar to the EuroSCORE, the STS-short term risk score was updated to adjust for changing patient characteristics.[23]

As this study shows, it is hard to determine when the time is right to update existing clinical prediction models. Ten years after the original EuroSCORE was reported to be no longer suitable for use due to decreasing performance.[5] This study shows that despite changes in the underlying population, at least in the Netherlands, the EuroSCORE II consistently overestimates mortality after cardiac surgery, but does not deteriorate. Other authors rightly point out that prediction models should not be deemed axiomatic,[23] yet leaving the question open when to update prediction models.

To assess prediction model deviations, other industries use the notion of concept drift.[2,24] Introduced to monitor sensor deviations in hardware industries, concept drift monitors how and when prediction models depart from their original outcomes (the 'concept'), or when one or more of their underlying predictors diverge.[24]

After detecting model drift, the models must be updated. Several authors have explored model updating of prediction models, in particular in cardiac surgery too.[19,25] To date, however, these dynamic modelling methods have not found their way to the models applied in clinical practice yet. Most likely, connections between clinical and methodological researchers need to be improved for that. Future research might focus on time limits for clinical prediction models to be (re-)assessed, and when needed, updated. Particularly when these models are greatly relied upon in clinical practice such as the EuroSCORE II.

Study limitations

Several limitations apply to this study. First, this study was performed using data from the NHR comprising a subset of Dutch cases. Regardless, the NHR is a very large and comprehensive registry. Similar analysis in other countries could therefore yield different results due to differences in populations and intervention criteria. Second, some predictors comprised up to 30% missing values, with less data missing as time progressed. Moreover, missing values were found to be not missing completely at random. Especially in such circumstances a complete case analysis, which excludes patients with one or more missing predictor or outcome values from the analysis, is known to lead to biased results.[26] Therefore, we feel confident that our approach was the best possible way to deal with the encountered problem of missing data.

CONCLUSIONS

Even though the population characteristics of cardiac surgery patients in the Netherlands have changed over time, the EuroSCORE II consistently overestimates the in-hospital mortality risk after cardiothoracic surgery in the Netherlands over time for all major procedures. Still, more attention is needed for dynamic trends and modelling in clinical prediction models.

REFERENCES

1. Nashef SAM, Roques F, Sharples LD, et al. Euroscore II. *Eur J Cardio-thoracic Surg*. 2012;41(4):734-745.
2. Kukar M. Drifting concepts as hidden factors in clinical studies. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2003;2780 LNAI:355-364.
3. Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardio-thoracic Surg*. 1999;16(1):9-13.
4. Siregar S, Groenwold RHH, de Heer F, Bots ML, van der Graaf Y, van Herwerden LA. Performance of the original EuroSCORE. *Eur J Cardio-thoracic Surg*. 2012;41(4):746-754.
5. Hickey GL, Grant SW, Murphy GJ, et al. Dynamic trends in cardiac surgery: Why the logistic euroscore is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur J Cardio-thoracic Surg*. 2013;43(6):1146-1152.
6. Nakatsugawa M, Cheng Z, Kiess A, et al. The Needs and Benefits of Continuous Model Updates on the Accuracy of RT-Induced Toxicity Prediction Models Within a Learning Health System. *Int J Radiat Oncol Biol Phys*. 2019;103(2):460-467.
7. van Veghel D, Marteijn M, de Mol B. First results of a national initiative to enable quality improvement of cardiovascular care by transparently reporting on patient-relevant outcomes. *Eur J Cardio-thoracic Surg*. 2016;49(6):1660-1669.
8. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. *Stat Med*. 2018;37(14):2252-2266.
9. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Oxford: John Wiley & Sons; 1987.
10. COX DR. Two further applications of a model for binary regression. *Biometrika*. 1958;45(3-4):562-565.
11. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Cham: Springer International Publishing; 2019.
12. R Core Team. R: A Language and Environment for Statistical Computing. 2020.
13. Velicki L, Cemerlic-Adjic N, Pavlovic K, et al. Clinical performance of the EuroSCORE II compared with the previous EuroSCORE Iterations. *Thorac Cardiovasc Surg*. 2014;62(4):288-297.
14. Margaryan R, Moscarelli M, Gasbarri T, et al. *EuroSCORE Performance in Minimally Invasive Cardiac Surgery Discrimination Ability and External Calibration*. Vol 12. http://www.rstudio.org/]. Accessed June 13, 2021.
15. Kuplay H, Bayer Erdoğan S, Baştopçu M, Karpuzoğlu E, Er H. Performance of the EuroSCORE II and the STS score for cardiac surgery in octogenarians. *Turkish J Thorac Cardiovasc Surg*. 2021;29(2):174-182.
16. Duchnowski P, Hryniewiecki T, Kuśmierczyk M, Szymanski P. Performance of the EuroSCORE II and the society of thoracic surgeons score in patients undergoing aortic valve replacement for aortic stenosis. *J Thorac Dis*. 2019;11(5):2076-2081.
17. Sullivan PG, Wallach JD, Ioannidis JPA. Meta-Analysis Comparing Established Risk Prediction Models (EuroSCORE II, STS Score, and ACEF Score) for Perioperative Mortality During Cardiac Surgery. *Am J Cardiol*. 2016;118(10):1574-1582.
18. Hogervorst EK, Rosseel PMJ, van de Watering LMG, et al. Prospective validation of the EuroSCORE II risk model in a single Dutch cardiac surgery centre. *Netherlands Hear J*. 2018;26(11):540-551.
19. Siregar S, Nieboer D, Vergouwe Y, et al. Improved Prediction by Dynamic Modeling: An Exploratory Study in the Adult Cardiac Surgery Database of the Netherlands Association for Cardio-Thoracic Surgery. *Circ Cardiovasc Qual Outcomes*. 2016;9(2):171-181.
20. Bedeir K, Kaneko T, Aranki S. Current and evolving strategies in the management of severe mitral annular calcification. *J Thorac Cardiovasc Surg*. 2019;157(2):555-566.

21. Kaneko T, Hirji S, Percy E, et al. Characterizing Risks Associated With Mitral Annular Calcification in Mitral Valve Replacement. In: *Annals of Thoracic Surgery*. Vol 108. Elsevier USA; 2019:1761-1767.

22. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):1-7.

23. Shahian DM, Jacobs JP, Badhwar V, et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1-Background, Design Considerations, and Model Development. 2018.

24. Gama J, Zliobaite I, Bifet A, Pechenizikiy M, Bouchachia A. A Survey on Concept Drift Adaptation. *ACM Comput Surv*. 2014;46(4):258-264.

25. Vergouwe Y, Nieboer D, Oostenbrink R, et al. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med*. 2017;36(28):4529-4539.

26. Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol*. 2006;59(10):1092-1101.

SUPPLEMENTARY MATERIALS

Table S1: Missing data per EuroSCORE II predictor in cardiac surgery within the Netherlands Heart Registration

Characteristic	**Patients with at least 1 Missing value (n = 36,947)**	**Patients with complete data (n = 66,457)**	**p-value**
Outcome			
In-hospital mortality:	1221 (3.37%)	1535 (2.31%)	<0.001
Predictor			
Age (years)*:	66.0 (11.0)	65.8 (11.0)	0.016
Female*:	10545 (28.5%)	17909 (26.9%)	<0.001
Weight (kg):	82.6 (15.2)	83.0 (15.3)	<0.001
Creatinine concentration (mmol/L)*:	93.1 (45.6)	93.1 (48.2)	0.808
Dialysis:	93 (0.52%)	321 (0.48%)	0.554
Diabetes mellitus:			0.000
No diabetes mellitus	24899 (78.2%)	52549 (79.1%)	
Diabetes mellitus, treatment unknown	2293 (7.20%)	551 (0.83%)	
Diabetes mellitus, no treatment	208 (0.65%)	441 (0.66%)	
Diabetes mellitus, diet	108 (0.34%)	535 (0.81%)	
Diabetes mellitus, oral medication	2644 (8.30%)	6824 (10.3%)	
Diabetes mellitus, insulin-dependent*	1598 (5.02%)	4782 (7.20%)	
Diabetes mellitus, other	94 (0.30%)	775 (1.17%)	
NYHA*:			<0.001
Class I	2427 (42.5%)	20561 (30.9%)	
Class II	1461 (25.6%)	24802 (37.3%)	
Class III	1488 (26.0%)	17308 (26.0%)	
Class IV	341 (5.96%)	3786 (5.70%)	
LV-function (%):	51.5 (9.20)	51.4 (9.61)	0.138
CCS4*:	737 (4.17%)	4072 (6.13%)	<0.001
Active endocarditis*:	778 (2.16%)	1698 (2.56%)	<0.001
Extracardiac arteriopathy*:	3929 (10.9%)	6937 (10.4%)	0.025
Chronic pulmonary disease*:	3957 (11.0%)	6821 (10.3%)	0.001
Pulmonary artery systolic pressure (mmHg):	17.2 (5.30)	17.9 (7.00)	<0.001
Recent myocardial infarction*:	6687 (18.5%)	13151 (19.8%)	<0.001
N/M mobility*:	261 (1.62%)	2990 (4.50%)	<0.001
Urgency*:			<0.001
Elective	22458 (70.4%)	41193 (62.0%)	
Urgent	6596 (20.7%)	20961 (31.5%)	
Emergency	2234 (7.00%)	3501 (5.27%)	
Salvage	614 (1.92%)	802 (1.21%)	
Critical preoperative state*:	261 (1.62%)	2990 (4.50%)	<0.001
Weight of procedure*:			<0.001
Isolated CABG	11827 (50.1%)	34109 (51.3%)	
1 procedure (non-CABG)	6210 (26.3%)	16108 (24.2%)	
2 procedures	4339 (18.4%)	12520 (18.8%)	
3 or more procedures	1219 (5.17%)	3720 (5.60%)	
Thoracic aorta surgery*:	3699 (10.2%)	5145 (7.74%)	<0.001
Previous cardiac surgery (redo) *:	2891 (7.99%)	4503 (6.78%)	<0.001

*: used in EuroSCORE II; CCS4: CCS class 4 angina; LV function: left ventricular function; N/M mobility: neurological or musculoskeletal dysfunction severely affecting mobility; NYHA: NYHA: New York Heart Association heart failure category; CABG: coronary artery bypass grafting. P-values: chi-square tests and t-tests were performed for categorical and continuous variables respectively.

Figure S1: Missing data per EuroSCORE II predictor over time

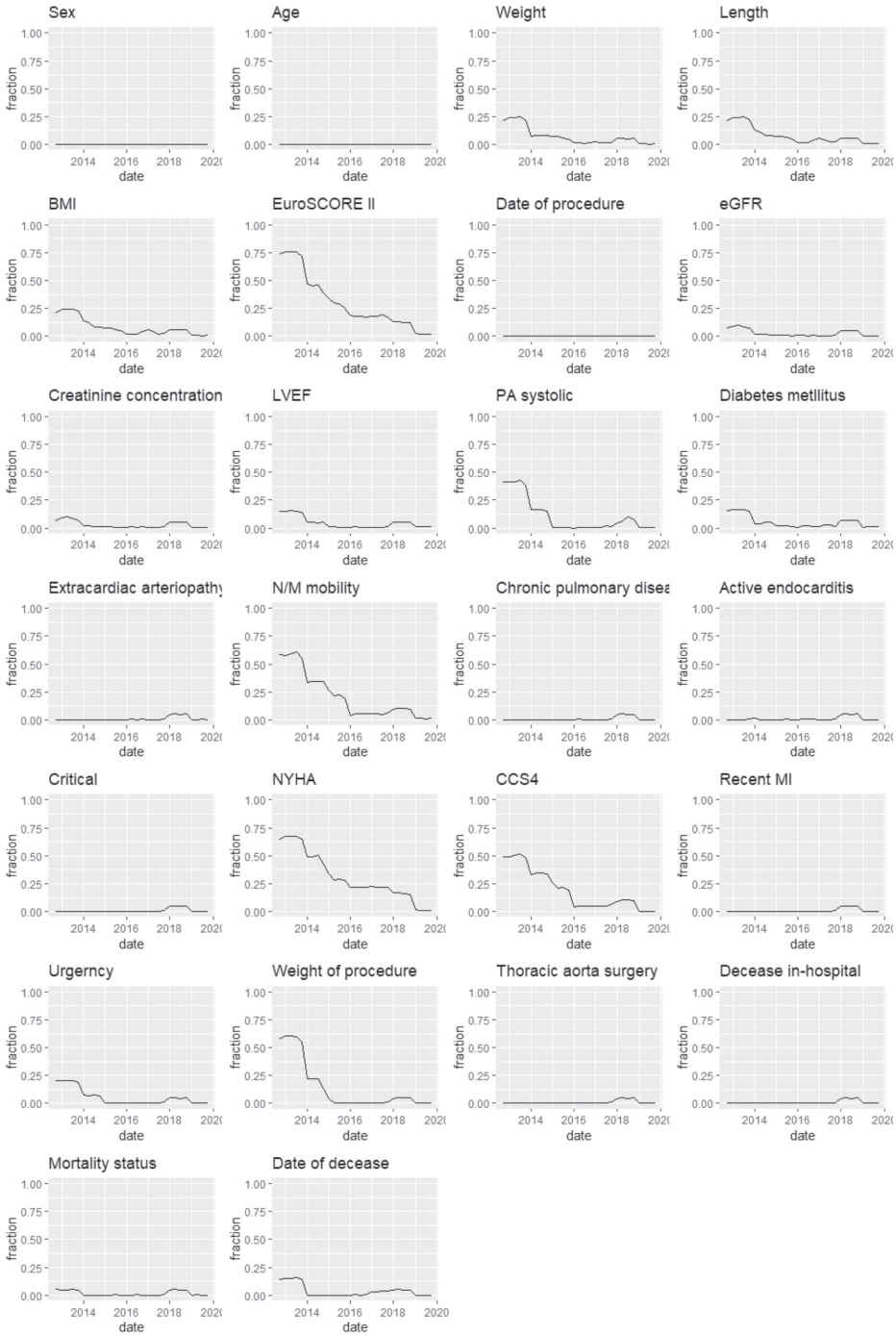
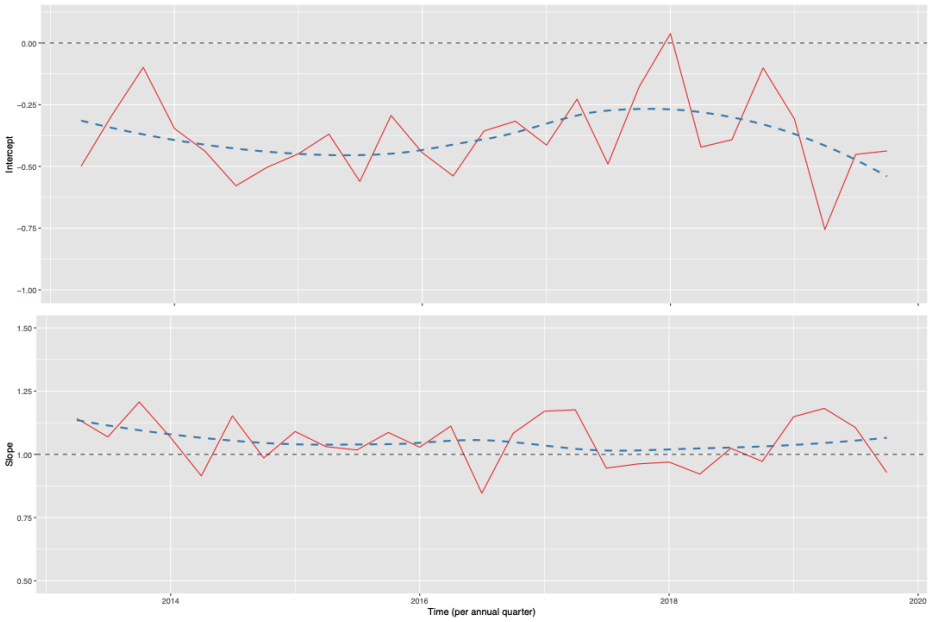
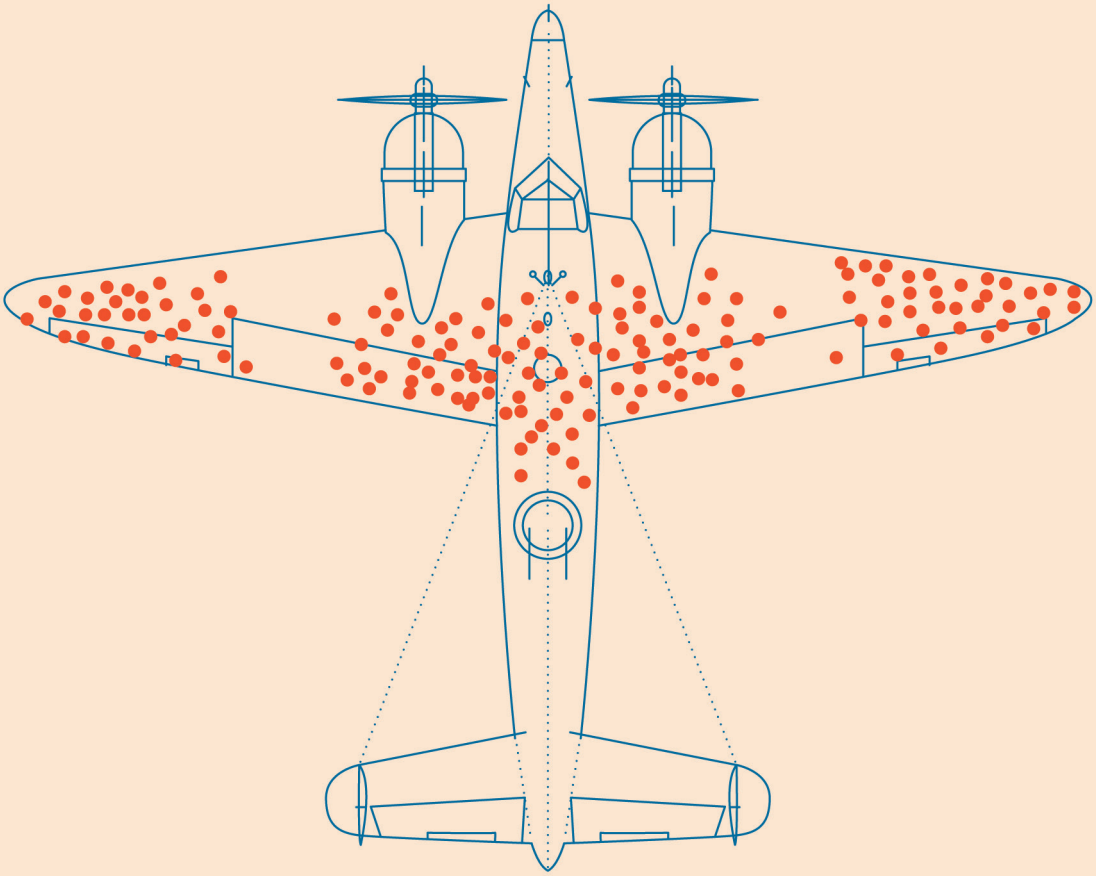


Figure S2: Calibration-in-the-large and slope of the EuroSCORE II over time



Legend, red line: CITL/slope; blue dotted line: trend line (LOESS) of CITL/slope.



Chapter 10

A guide to privacy-preserving record linkage

Wouter B van Dijk, Mira G P Zuidgeest, Ewoud Schuit, Rieke van der Graaf, Menno Mostert, Julie Anne R Smit, Martine C de Vries, Diederick E Grobbee, Ilonca Vaartjes

Submitted.

ABSTRACT

Background

Combining datasets is of high importance as it increases the learnings for health care and society, but preservation of privacy can be a challenge. Privacy-preserving record linkage (PPRL) methods can assist in combining data while protecting privacy of individuals.

Objective

To provide guidance for researchers in PPRL by offering an overview of available methods and a step-by-step approach to PPRL.

Study Design and Setting

A systematic review on PPRL was performed from which key steps and patterns were extracted and translated into a step-by-step PPRL approach.

Results

We propose a five-step approach to perform PPRL. First, to assess datasets on their privacy risk level. Second, to determine the applicable party collaboration model. Third, to select an adequate PPRL method. Fourth, to determine whether to use additional privacy enhancing methods. Finally, to assess the quality of their chosen approach.

Conclusion

We provide a structured approach to help researchers create transparent and high quality PPRL.

INTRODUCTION

Record linkage aims to combine two or more datasets into a single, enriched dataset to facilitate health research with new valuable insights that would not have been possible when analyzing the datasets separately.[1,2] Traditionally, guidance on record linkage focusses on resolving conflicts between datasets, also known as entity resolution, either by adding columns-or by adding and finding duplicates.[3] However, protecting subject privacy can be difficult or even impossible when using these methods.[4] Common linkage methods often rely on linking subjects on a combination of their forename, last name, date of birth and sex or race, inherently limiting privacy protection.[3,4]

With increasing use of record linkage resulting from big real-world data research, data protection regulations expect record linkage methods to do more to preserve subject privacy, i.e. prevent subject identity disclosure.[5,6] For example, the European General Data Protection Regulation (GDPR) warrants researchers to use *privacy by design*-methods and perform Data Privacy Impact Assessments (DPIAs) in their records linkage processes.[7] Moreover, the need for more advanced methods to protect subject privacy was recently shown again by the US Census Bureau, reporting that 46 to 71% of their records could theoretically be matched to reidentify subjects when combined with data publicly available.[8]

To overcome challenges of traditional record linkage methods, specific methods facilitating Privacy-Preserving Record Linkage (PPRL) for biomedical research have been developed.[4] For users, however, it is hard to get an overview on PPRL methods, as current literature mainly concerns individual methods or integrated system solutions.[9] Current literature generally does not address which method is best suited in which situations. The two overviews that do exist are limited as they only list available methods, while any guidance on how to arrive at solid PPRL still lacks.[4,10] As a result, it is often hard for new entrants to determine which factors are important to take into account and what steps to consider to come to reliable PPRL.

The aim of this guidance paper is to assist researchers in reliably linking datasets whilst protecting subject privacy. We provide a guide on PPRL presented as a step-by-step approach describing factors researchers need to take into account when they want to link records, while aiming to protect subject privacy. Our approach is substantiated on a

preceding systematic review aimed at the identification of PPRL methods used in medical practice and research (see **Supplementary Materials** for details on the systematic review).

Besides helping researchers to link data reliably, we would like to encourage them to also report on the steps they have taken in methods sections of their paper as to improve reporting on PPRL.

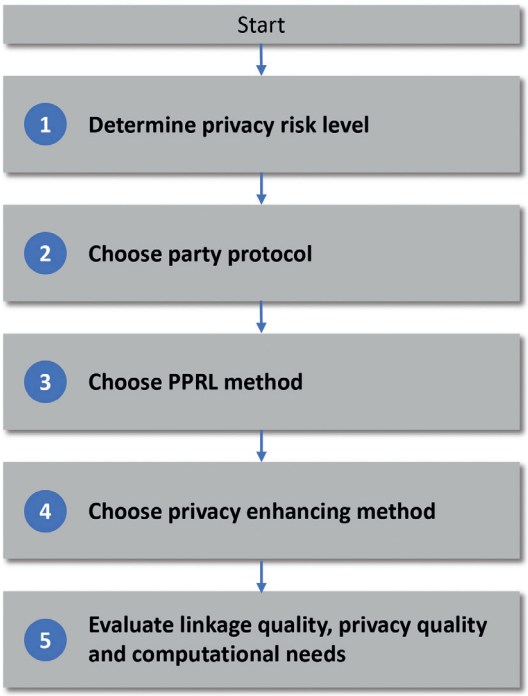
STEPS OF PRIVACY-PRESERVING RECORD LINKAGE

Five main steps for PPRL were identified (**Figure 1**) from the preceding systematic review described in the **Supplemental materials**. From included literature common themes and methods were extracted by categorizing articles on the main subject addressed in their main text. For the current guide, only methods combining participant privacy protection and record linkage were included.

First step in the PPRL process is assessing the privacy risk levels of the individual datasets to link and those of the created dataset after record linkage. Second step in the PPRL process is to choose a party protocol. Party protocols describe how parties involved in a PPRL case want to collaborate and share data with one another. The different party protocols also reflect the level of trust between parties. Third step is selecting the PPRL method most appropriate for the given situation. Important to take into account is that every record linkage depends on a trade-off between linkage quality and the protection of subject privacy.[8] This trade-off is primarily reflected in the method chosen to apply. Fourth step is choosing possible additional privacy protection methods to further protect subject privacy before or after linkage has been conducted. Final, and fifth, step is to review the quality, privacy and computational needs of the linkage process resulting from choices made in the previous steps.

Identified steps are discussed hereafter. Each step is illustrated with an example figurating a Dutch cardiovascular linkage case.

Figure 1: Flow chart



Box 1: Use case, drug treatment efficacy for secondary prevention of cardiovascular disease

We take a hypothetical use case wherein the aim is to compare the efficacy of ACE-inhibitors in heart failure patients with various ethnicities. A cardiovascular disease registry comprising routinely collected healthcare data is used as base and linked with the national population and death registry. The disease registry mirrors the Utrecht Cardiovascular Cohort, comprising patient characteristics as age, gender, heart failure stage, treatments received, comorbidities and potential confounding risk factors like BMI, blood pressure, smoking status and cholesterol of 500.000 patients. Linkage with the national population and death registry is required as ethnicity, socio-economic status and cause of death are not part of the disease registry.

Step 1. Determine privacy risk levels

Before linking any records with PPRL, involved parties should carefully analyze benefits, risks and available safeguards based on the context in which records will be linked.[9,11,12] The goal in record linkage is to exactly identify subjects and link information of multiple records.[4] The resulting dataset should then allow to make inferences on an individual level.[4]

Before linking data, it is important to distinguish two types of disclosure that can result from the linkage: *identity disclosure* and *attribute disclosure*.[4] Identity disclosure is when the identity of a participant is disclosed, he or she can then be reidentified from the data released. *Attribute disclosure* is the disclosure of one or more sensitive data attributes of participants like blood pressure, BMI or a genetic predisposition. Attribute disclosure can result in identity disclosure if a combination of attributes is unique in a dataset.[4,13] Awareness of risks of attribute disclosure is therefore important in PPRL. The aim of PPRL should be to find "the exact level of information disclosure that protects sensitive data but reveals enough identifying data for high quality linkage". Awareness on the privacy risks associated with the record linkage is therefore important.[13]

Factors often returning in literature as important for data privacy levels are: data granularity, sensitivity and precision.[14,15] Data granularity is the individuality of dataset subjects and attributes. For example, does a dataset contain individual subject data or are subjects grouped, and are attributes included separately, grouped or left out when irrelevant? Data sensitivity is the type of information in attributes. Attributes like HIV status and psychiatric history are generally regarded more sensitive than blood pressure and previous knee injury history. Data precision is the exactness of data, answered by the question: are values given exact or approximates?

To determine the privacy risks associated with record linkage we propose to create a privacy risk score for the information in datasets to be linked using a matrix with data granularity, sensitivity and precision as axes (**Figure 2**). As it is often hard to formally quantify risk levels for these factors, we developed a normative scoring system to grade information based on privacy risk factors most named in literature (**Table 1**). To reduce a privacy risk score researchers can group subjects or attributes (reducing data granularity), exclude sensible attributes (reducing data sensitivity), and obfuscate exact data values (reducing data precision).

Figure 2: Privacy risk levels

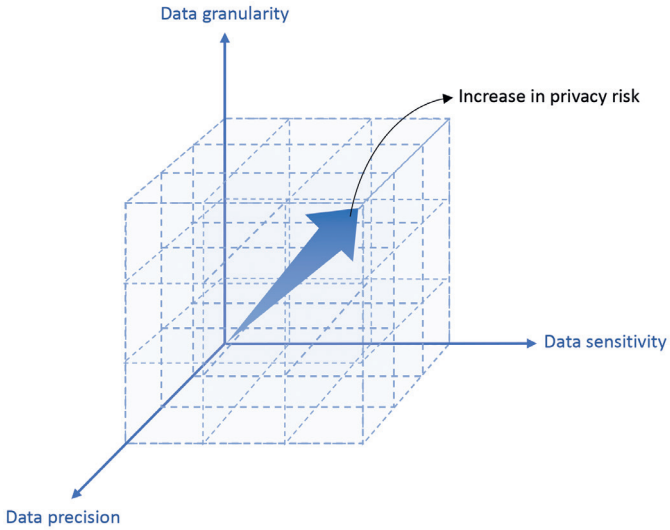


Table 1: Privacy risk assessment

Axis	Score	Description
Data granularity	1	Both, subject and attribute data are only linked on group or propensity score level
	2	Either subject or attribute data is grouped to preserve privacy
	3	Both, subject and attribute data are linked on individual level
Data sensitivity	1	None of the data linked is considered to be sensitive for subjects
	2	Some of the data linked is considered to be sensitive for subjects
	3	All of the data linked is considered to be sensitive for subjects
Data precision	1	All data attributes are approximate of their original values; no data attributes are linked on their original exact values
	2	Some data attributes are approximate of their original values; some data attributes are linked on their exact original values
	3	No data attributes are approximate of their original values; all data attributes are linked on their original exact values
Sum	...	

Interpretation of a calculated risk score will differ on a case-by-case basis, dependent on the type of research and question to be investigated. To protect oneself against justifying any risk score we advise to discuss a threshold risk score that is considered to be acceptable before calculating it. A more extensive example of privacy risk levels can be found in the article by Fernandes and colleagues, describing how sensitivity levels can be calculated for DNA reidentification risks.[14] More technical methods are described by

Kuzu and colleagues in their article on various mathematical methods on how to reduce privacy risk levels.[15]

Important to bear in mind is that when data are not grouped (i.e. reduced in granularity) a risk of deidentification will always remain.[12,15] Illustrative example of this was a deidentified dataset with user ratings released by Netflix in 2007 as part of a competition to compete against Netflix's predictive algorithms.[16] The data proved to be re-identifiable with auxiliary information from other sources. In this case, researchers were able to re-identify subjects by linking the data with IMDB user accounts. A study from 2017 showed that grouping attributes in propensity scores only harmed record linkage quality limitedly, offering additional safeguards options to maintain subject privacy.[17]

Use case

In our use case it would be understandable that our researchers initially planned to link their data on an individual subject level, comprising separate attributes. This would have resulted in a risk score of eight (granularity: 3, sensitivity: 2 [data containing ethnicity], precision: 3). When reassessing the data however, researchers could, for example, decide to calculate propensity scores for all attributes except heart failures stage, treatment received and ethnicity, resulting in a risk score reduction of two points (granularity: -1, precision: -1).

Step 2. Choose party protocol

Record linkage is commonly separated in three types of protocols based on the number of parties involved: two-party, three-party and multi-party (**Table 2**).[9,12,18] Two-party protocols concern direct record linkage between parties, in these protocols data owners and researchers are regarded as belonging to the same party. In addition, data protection can be added by allowing linkage in a secure environment only, controlled by one of the two parties. Three-party protocols regard indirect record linkage, where data is linked by a third-party. This party can be an independent researcher or a trusted-third party (TTP) that passes linkage keys or linked data to the researcher. Multi-party protocols regard linkage of more than two datasets, sometimes involving TTPs. Multi-party models often involve secure multi-party computations on which is elaborated later. As an alternative to data linkage, lately some researchers have been advocating that 'methods [could] travel instead of data' in the concept of a personal health train.[19] In a personal health train a data owner receives methods and algorithms to execute on their data. In this way a data owner can share the results of these methods without sharing the data. Personal health trains are considered outside the scope of this study and will thus not be discussed further.

Table 2: Party protocol models

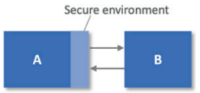
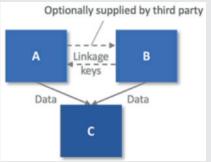
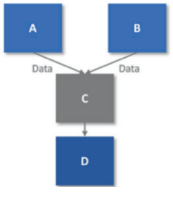
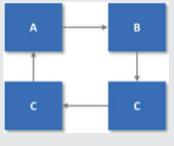
Parties involved	Collaboration form	Description	Advantages	Disadvantages
Two	*Direct linkage* 	Describes a situation in which two parties directly exchange data with one another. Researchers and institutions are regarded as one. This model is based on trust, optionally safeguarded by legal contracts, as both parties possess identifiable and privacy-preserved forms of linked datasets allowing privacy violations in case of malicious intent.	• Easiest model to achieve record linkage • Assumes equality between parties • Both parties to have advantage of linkage	• Least safe for PPRL
	Secure environment linkage 	Describes the situation in which one party offers a secure environment to link records. The party offering the environment will have control over data to show and at what granularity. A secure environment allows other parties that want to link to the data to query this data, allowing less rigid preparation constraints. This model is often combined with differential privacy and obfuscation features.	• Gives full control to offering party • Allows data exploration before linkage to some extend • Allows easy and dynamic differential privacy and obfuscation features	• Individual subject data linkage is impossible to preserve privacy • *Other parties will have to upload their data to the secure environment to link to data of the offering party*
Three	*Independent researcher linkage* 	Describes a situation similar to that of direct record linkage, yet with a researcher independent from the parties owning the data. In this situation none of the parties originally possessing the data will acquire the data from the other party. Optionally linkage keys can be exchanged between parties owning the data to deidentify data before issuing it to the researcher.	• No party with fully identifiable dataset when linkage keys are exchanged	• Risk for collusion with researcher • Researcher might be able to reidentify subjects using quasi-identifiers or frequency attacks • Entity resolution can be difficult if records are not properly prepared by parties owning data

Table 2: *(continued)*

Parties involved	Collaboration form	Description	Advantages	Disadvantages
	Trusted-third party (TTP) linkage 	Describes the situation in which a trusted third party performs the linkage and privacy preserving steps and forwards the privacy-preserved data to the researcher. In this way only the TTP will have all data. To lower privacy risks, parties can decide to first exchange linkage keys and share non identifiable data to the TTP afterwards.	• When linkage keys are exchanged beforehand no party will possess all identifiable data • Limits need for additional PPRL methods as privacy is preserved by TTP	• Risk for collusion with the researcher • Researcher might be able to reidentify subjects using quasi-identifiers or frequency attacks • Parties will heavily rely on capabilities of TTP on linkage and on privacy-preserving measurements
Multi	*Aggregate multi-party linkage* 	Describes the situation in which more than two parties want to exchange data, but do not require individual subject data (if linking individual subject data is desired between more than two parties one of the previous models can be used). In this situation aggregates are shared between parties after each party has manipulated the data in a predetermined way. Manipulation per party can then be kept secret while the sum of the manipulations of all parties is released. In this way parties can exchange data and derive insights without sharing individual level data.	• Limits need for additional PPRL methods as privacy is preserved by computations	• No individual subject data linkage possible

Next to the *de facto* number of parties involved in the PPRL process, researchers are advised to take three adversary models into account, reflecting the amount of trust between parties.[3,20–22] The following three adversary models are described in literature:

1. *Honest-but-curious (HBC)* – assumes that parties follow a given protocol while still being curious to learn more about the other party's data. This model is based on trust and is assumed by most PPRL methods.[4] Previous studies have shown that an attack on privacy often comes from insiders in an HBC model, mostly found in unauthorized access to and use of information (63%) and unintentional privacy breaches (57%).[4]
2. *Accountable computing/covert* – assumes parties will adhere to protocol by registering protocol interactions. This requires the creation of accountability measures besides linkage and privacy-preserving measures.[23]
3. *Malicious* – assumes that parties will behave arbitrarily, breaking protocol when it is to their advantage. Protocols under this model either need to use a TTP or highly complex PPRL methods.

Use case

In our use case only two parties *would be* involved. The institution where the disease registry is hosted and the controller of the national population and death registry. As adversary model parties *could reasonably* settle on *an* accountable model. Data linkage *would* be executed in a secure environment controlled by the controller of the national registry. The secure environment *would allow* parties to link freely while preventing data leakage, monitoring user activity to protect subject privacy.

Strep 3. Choose PPRL method

In literature on PPRL, record linkage methods, privacy preserving methods and methods combining preserving privacy with record linkage ('real' PPRL) are used interchangeably.

Most PPRL methods make direct or indirect use of attributes comprising subject identifiers (IDs) of datasets to link (**Table 3**). Methods using direct subject IDs do this by replacing attributes on which subjects can directly be identified with random keys or encrypted substitutes. These keys are also known as statistical linkage keys (SLKs).[22,24] Identifying keys can be created using single fields or by combining multiple fields.[10] When using a TTP, an elegant method is to just remove all direct subject IDs, returning the data without identifiers (decoupled data) after linkage.[4,25] Because the mentioned methods rely on direct subject IDs they are easy to use. Yet, these methods generally leave indirect subject IDs unprotected, making data vulnerable to reidentification.[9]

Bloom filters are considered as the reference standard for record linkage by many, because they allow combining privacy preserving features with approximate linkage.[9,26,27] Bloom filters work by translating subject IDs into cryptographic keys. They do this by splitting IDs into bigrams (e.g. Smith will be split into '_S', 'Sm', 'mi', 'it', 'th' and 'h_'), which are then hashed and, subsequently, translated into bit arrays (e.g. 110000100). These bit arrays can be assessed on similarity without disclosing a subject's ID. An comprehensive description of Bloom filters can be found in the original article of Schnell and colleagues.[28] Furthermore, many R scripts exist on the internet and can be easily found.

Although often considered as de facto standard for PPRL, Bloom filters have been shown to be prone to decoding attacks. [29] In decoding attacks the original data is reconstructed based on patterns in the cryptographic keys, making it possible to reidentify original values. Several modifications to Bloom filters have therefore been proposed to reduces these vulnerabilities. [30–33] These additional privacy features, however, increase computational power needed to calculate the cryptographic keys, introducing different challenges for large datasets.[30] In addition, a variance on Bloom filters has been developed to secure numerical data as original Bloom filters are only aimed at securing and linking string data.[34]

A different approach for PPRL is garbled circuits. In garbled circuits attributes are encoded in a logic circuit.[21] One party creates the logic circuit, taking into account potential inputs of the other party. The other party can then evaluate the circuit by using its input. Explained more easily, this can be compared to a situation in which party one hands over a key to party two, on which the latter party then generates a lock unique to that key. Disadvantage of this method is that it assumes an unequal relation between parties in which one party is forced to expose its inputs to allow the other party to create the lock. Moreover, garbled circuits are not scalable, one will need to generate a garbled circuit for every time datasets are linked.

Last are secure multi-party computations (SMCs). SMCs are a secure approach to data privacy yet are only suited for aggregate data sharing and do comprise executional challenges as they require all parties to perform operations on shared data.[9,12] In SMCs each party computes part of a function while keeping their individual part private (e.g. s = 99 [random starter] + 11 [party 1] + 23 [party 2] + 17 [party 3]). After the computations are finished, each party knows the end result and their own part, enabling them to derive the sum of data of the other parties (deriving the sum of other party data for party 1: s = 139 [sum of random start, parties 2 and 3] – 11 [own part] – 99 [random starter] = 29).

When exchanging aggregate data between three or more parties we would also like to point at other methods than mentioned in this paper like meta-analytical, stratified tabular and distributed regression to avoid the need of PPRL.[17] Since our focus was on linkage of individual level data, linkage on aggregate level data was considered outside the scope of this study.

Use case

As data would only be shared within a secured environment where all operations were monitored in our use case, it could be decided to link records based on (simple) encrypted statistical linkage keys (SLKs). The SLKs could easily be created by performing a standard SHA-256 hash combination over key identifiers not prone to errors (date of birth, sex, zip code and house number). SHA-256 algorithms are supported standard by all major statistical software packages, include R base. Advantages of using this method would be computational speed and accuracy. Disadvantages would be higher numbers of missed record linkages (as keys must match exactly) and lower quality of privacy protection when attacked, both risks could be considered acceptable for the case at hand considering that the data will not leave the secure environment and the linkage process would be actively monitored.

Table 3: PPRL methods

Party protocol	Method	Linkage method	Description	Advantages	Disadvantages	Linkage capa-bilities	Privacy capa-bilities	Compu-tational needs	Source
>=2	Random linkage keys	Exact matching (deterministic)	Remove identifiable information and replace these with a randomly generated linkage key. Linkage keys can be supplied by a third party (e.g., U.S. National Institutes of Health Global Unique Identifier (GUID) or European patient-identity management solution (EUPID)) or a computer algorithm (e.g. Mainzelliste)	Easy to execute	Subjects can often still be identified on unique combinations of attributes Datasets need to be harmonized before generating linkage keys Vulnerable for frequency attacks	+/-	+/-	+	Kum 2014; Baker 2019; Canterberry 2019; Durham 2012; Christen 2016; Stammler 2020; Vatsalan 2013; Randall 2019
	(Secure) hash-linkage key /hash-encoding	Exact matching (deterministic)	One way hash function like SHA to "encode-and-compare" data. Can be used to create unique linkage keys from IDs to link datasets upon or to hide data attributes.	Easy to execute	Subjects can often still be identified on unique combinations of attributes Single character difference between datasets will result in difference hashes Vulnerable for frequency attacks Vulnerable for dictionary attacks	-*	+/-	+	Christen 2016; Durham 2012; Baker 2019; Bian 2019; Agiro 2019; Guesdon 2016; Vatsalan 2013
	Encrypted statistical linkage key (ESLK)	Exact matching (deterministic)	Linkage key has with one-way hashing derived from components of ID attributes (e.g. family name + given name + date of birth becomes XXXZZYYYMMDD).	Relatively easy to execute	Limited privacy protection Vulnerable for frequency attacks	+/-	+	+	Baker 2019; Agiro 2019; Bian 2019; Chen 2018; Schnell 2018

Table 3: *(continued)*

Party protocol	Method	Linkage method	Description	Advantages	Disadvantages	Linkage capa-bilities	Privacy capa-bilities	Compu-tational needs	Source
	(Field level) Bloom filter encoding	Approximate matching (probabilistic)	Bit arrays of n-grams of IDs into Bloom filters resulting in encoded hashes through which data can be linked on approximation. Traditional Bloom filters were developed to encode strings, today Bloom filters also exist for numerical data.	Allows approximate data linkage	Vulnerable to cryptographic/decoding attacks	+	-	+/-	Durham 2012; Baker 2019; Demelius 2020; Brown 2019; Brown 2017; Schnell 2009; Vatsalan 2013
	(Record level) composite Bloom filters / crypto-graphic long-term keys	Approximate matching (probabilistic)	Composite Bloom filters are Bloom filters based on multiple ID components to enhance privacy protection.	Allows approximate data linkage Stronger protection against cryptographic attacks	Still vulnerable for cryptographic attacks (yet less than traditional Bloom filters)	+	+/-	+/-	Brown 2017
	Markov Chain-based Bloom Filter	Approximate matching (probabilistic)	Markov Chain-based Bloom filters are also Bloom filters yet using Markov Chains to enhance privacy protection.	Allows approximate data linkage Stronger protection against cryptographic attacks	Computationally expensive	+	+	-	Schnell 2018
	Garbled circuits	Exact matching (deterministic)	Garbled circuits are based on logic circuits in which one user create a 'lock' around a 'key' create from the data of the other party	Secure, precise data linkage	Unequal relation between parties Not scalable, requires a new circuit for each linkage	+	+	-	Chen 2018

Table 3: *(continued)*

Party protocol	Method	Linkage method	Description	Advantages	Disadvantages	Linkage capa-bilities	Privacy capa-bilities	Compu-tational needs	Source
>=2 w/ TTP	Decoup-ling data	Exact matching (deterministic)	Dataset owners send data comprising identifiable data to TTP which links data and subsequently removes ID's	Traditional record linkage can be used to link datasets	Requires TTP Datasets needs to comprise similar identifiers to allow linkage Vulnerable for collusion	+	+/-	+	Kum 2014; Boyens 2004
>=3	Multi-party Bloom filters	Approximate matching (probabilistic)	Multi-party Bloom filters allow approximate matching of records between more than two parties	Allow record linkage between more than two parties	Vulnerable for cryptographic/decoding attacks	+	-	+/-	Christen 2016; Vatsalan 2013
	Encryption schemes / secure multi-party compu-tation (SMC)	Exact matching (deterministic)	Computes a function across several parties, such that no party learns the information from the other parties, but all receive the final results. Most used: - Secure scalar product - Secure set intersection - Secure set union	Allows data linkage with precision while keeping individual party data secret	Computationally expensive	+/-	+	+/-	Christen 2016; Baker 2019; Kuzu 2013; Laud 2018; Lu 2015; Vatsalan 2013

* Single character differences will result in un-linkable data

Step 4. Choose privacy enhancement method

Besides linking data using a PPRL method, additional privacy enhancements can be added to protect subject privacy in linked datasets (**Table 4**). These additional privacy enhancements can either be applied before linkage, limiting release of data, or after linkage. Adding privacy enhancements before linkage is used to protect subject privacy against reidentification by a party with whose data will be linked. Adding privacy enhancements after linkage is used to protect the combined dataset against reidentification by third-party privacy enhancement methods provide additional privacy protection by reducing data granularity and/or precision.

Additional privacy enhancement can be applied in two ways: by value generalization and by obfuscating values.[4,35] Value generalization methods focus on limiting the amount of data released by grouping data. Most famous herein is the differential privacy method.[4,8,12] Differential privacy relies on a privacy-loss budget, telling how much additional information can be inferred from adding or deleting a record in a dataset, practically giving the chance of identifying an individual subject from the data released.[8] By choosing a reidentification threshold for subjects in the form of a privacy-loss parameter, data owners can ensure additional privacy protection to data released. An Important feature of differential privacy is that it also allows to quantify how many queries can be executed before subject privacy is threatened by calculating combined reidentification chances.[8] Extensive descriptions of pros and cons of using differential privacy can be found in the report substantiating the choice of the US Census Bureau to adopt differential privacy as primary subject protection method from 2020 onwards.[8] Implementing differential privacy methods can look daunting at first, fortunately guides[36] and R-packages, like '*diffpriv*',[37] are available facilitating differential privacy today.

Other value generalization methods that aggregate subject data include k-anonymity, sensitivity level binning and clustering based anonymizing algorithms.[12,14,38] Another option is propensity score summarization, which differs from the other value generalization methods by grouping attribute data instead of subject data. An extensive description of this method can be found in a paper of Bohn and colleagues.[17,39]

Value obfuscation methods operate by hiding original exact values.[4,40] Methods described hide original values by introducing random or systematic errors on an individual level while maintaining overall summary statistics.[4,12,40] Moreover, original values can be hidden by adding fake data (noise) to the dataset, again keeping summary statistics similar.[4,12]

Reference space embedding is a method in which a randomly generated set of strings from a reference set are embedded in the data, adding space to it.[12,40] The distance added by the strings can be shared between parties, allowing to compare data on similarity when taking the given space into account.

Use case

In our use case, adding k-anonymity to the secure environment as additional privacy preserving layer would be conceivable. K-anonymity privacy protection approximates differential privacy in that it tries to determine the risk of subject identification, yet does this in a less formal matter, only testing subject identification on prespecified parameters against a static denominator. Although k-anonymity is criticized on being imperfect also, it would provide easy to implement additional protection. Moreover, re-identification risks through k-anonymity could be actively monitored in the secure environment.

Table 4: Privacy enhancing methods

Type	Method	Description	Source
Value generalization	Differential privacy	The probability of holding any property on the perturbed database is approximately the same whether or not an individual value is present in the database.	Christen 2016; Kum 2014
	K-anonymity	K-anonymity works on denominator defining the reidentification chance of data released. When the number of reliabilities is smaller than the denominator (i.e. data from less subjects than defined is necessary to reidentify subjects) data is not released.	Christen 2016
	Sensitivity level binning	Prevent rare combinations to be found. Sensitivity levels based on frequencies are built on the fact that a rare disease/genetic variation should be considered more sensitive than a common one, since they concern a smaller subset of the population.	Fernandes 2020
	Clustering-based anonymizing algorithms	Algorithm based anonymization that automatically selects sets of records to be clustered.	Loukides 2010
	Propensity score summarization	In propensity score summarization data attributes of subjects are combined into a single propensity score to keep individual attributes secret.	Bohn 2017; Guesdon 2016
Value obfuscation	Error introduction	Random or systematic addition or subtraction of values of data points to reduce data precision.	Durham 2012
	Noise addition / chaffing	Random or systematic addition or deletion of rows or data points, while keeping means, medians, and standard deviation consistent.	Christen 2016; Kum 2014
	Reference space embedding	Base values on distance between attributes Public or private.	Christen 2016; Kum 2014; Durham 2012

Step 5. Evaluate linkage quality, privacy quality and computational needs

Many authors advise researchers to test PPRL on the following three features: 1) linkage quality, 2) privacy quality and 3) computational needs.[3,4,22,40] Methods strong on one feature often pay a price on performance on another feature.[40]

Linkage quality is generally assessed by the true positive rate (defined as true positives / [true positives + false positives]).[40] Other usable measures include diagnostic performance measures as sensitivity, specificity and the F1-score.[41] Privacy quality is measured in the mutual information entropy, entropy information gain or relative information gain. Comparable to a privacy-loss budget entropy represents the amount of information that can be acquired from data, being a derivative of the risk of reidentifcation.[3,40] Calculating entropy scores is standard practice in security evaluation to quantify the relation between the original fields and encoding.[40] Examples of entropy calculation have been published by Airoldi and colleagues.[42] Also, Wagner and colleagues published a review of technical privacy metrics discussing technical details of the entropy method in depth.[43] The proof of privacy quality can be evaluated by simulating several solutions under different adversary models as described by Durham and Wagner.[3,40,43] Again, R-packages to support calculating entropy scores, like '*EntropyExplorer*'[44], are available online. Computational needs are simply measured as the runtime a PPRL method needs for performing its tasks.[40]

How to interpret and evaluate found PPRL performances metrics will depend on the circumstances per PPRL situation. For support, several authors do give suggestions based on theory or use cases.[40,42,43]

Use case

When statistical linkage keys would be chosen to protect subject privacy a true positive rate of 60 to 100% could be expected from the linkage, depending on the similarity and quality of the individual datasets.[31] Computational runtime could be assumed to be around 300 seconds per 10^6 records[40], resulting in an expected runtime of 150 seconds when processing 500.000 records. Entropy could be calculated as additional performance parameter, but as data protection in our use case would result from the secure environment with monitoring instead of the quality of SLKs this could also be left out in our use case.

DISCUSSION

In this article we provided a step-by-step process to guide researchers through their PPRL process. We identified five steps to be taken subsequently to determine which PPRL approach to take for a specific linkage situation. To be able to develop a rigorous method, a systematic review was performed in advance. Moreover, we separated methods described in literature in linkage, privacy and combined PPRL methods to be able give clearer guidance on when to use which methods.

Previously, PPRL methods have been compared.[3,4,40,45,46] Important for visibility of PPRL methods was the article by Vatsalan et al., who first described available PPRL methods. Vatsalan et al. also provided a detailed overview of technical methods available for PPRL and is unique to date in providing such an overview. Durham et al. provided an overview on PPRL methods by comparing methods on their performance on linkage quality, privacy quality and computational needs. No articles, however provided a high-level overview reporting on the consecutive steps researchers should take to reach PPRL. We developed a literature-based stepwise guide to support researchers with PPRL.

The challenges of developing high quality PPRL methods are demonstrated in the fact that since PPRL were mentioned first in the ’70s and ‘80s only a limited number of methods have been established. Bloom filters as introduced by Schnell and colleagues in 2009 are the most recent innovation in PPRL and were swiftly adopted. Before Bloom filters were introduced, limited quality of privacy protection in record linkage was often accepted as no other methods allowed approximate record linkage with preservation of privacy until then. In recent years multiple authors, have taken initiative to improve privacy preserving features of Bloom filters in order to tackle their biggest weakness.[30,47,48] Adequate alternatives for Bloom filters, especially methods that allow for probabilistic linkage, however, are in dire need still.

Moreover, broader acceptation that reidentification is always, to some extent, at odds with accuracy of data is important. Major step forward was the embrace of the United States Census Bureau, that announced to start using privacy budgets to decide on what data to release and on what level in their 2019 operational plan. With this announcement the US Census Bureau acknowledged that privacy will always come at the cost of data accuracy, i.e. increased subject privacy protection is inversely related to data accuracy.[49]

Study limitations

A couple of limitations apply to this article.

First, this guide could only be developed after structuring and interpreting literature found in the related systematic review. As a consequence, it does not only represent literature available on PPRL, but also the interpretations of its authors. To minimize author preferences, conflicts were discussed for consensus, yet remain a reflection of this group of authors.

Second, to provide an overview of PPRL methods currently used in medical research, the systematic review substantiating this article was performed in biomedical databases only. As a result, novel PPRL methods explored in the field of computer science that might be of value in the biomedical context might have been missed.

Third, no details on specific PPRL methods were described in this article. Nuances of individual PPRL methods might therefore be underexposed in this article. For more critical reviews of methods mentioned we refer to specific papers mentioned, and other references attached to this article.

CONCLUSIONS

After literature search five steps were identified for researchers to follow when performing PPRL. This systematic stepped approach could help to guide researchers through a PPRL process and to improve the correct application of PPRL methods in data research.

REFERENCES

73. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. Science (80). 1959; 130(3381):954-959.
74. Gu L, Baxter R. Record linkage: Current practice and future directions. Cmis. 2003:03/83.
75. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. Inf Syst. 2013;38(6):946-969.
76. Kum H-C, Krishnamurthy A, Machanavajjhala A, Reiter MK, Ahalt S. Privacy preserving interactive record linkage (PPIRL). J Am Med Informatics Assoc. 2014;21(2):212-220.
77. Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. Nat Med. 2020;26(1):29-38.
78. Harron K, Dibben C, Boyd J, et al. Challenges in administrative data linkage for research. Big Data Soc. 2017;4(2):1-12.
79. Kalkman S, Mostert M, Gerlinger C, Van Delden JJM, Van Thiel GJMW. Responsible data sharing in international health research: A systematic review of principles and norms. BMC Med Ethics. 2019;20(1):1-13.
80. The MITRE Corporation, US Census Bureau. Formal Privacy Methods for the 2020 Census.
81. Baker DB, Knoppers BM, Phillips M, et al. Privacy-Preserving Linkage of Genomic and Clinical Data Sets. IEEE/ACM Trans Comput Biol Bioinforma. 2019;16(4):1342-1348.
82. Vatsalan D, Christen P. Privacy-preserving matching of similar patients. J Biomed Inform. 2016;59:285-298.
83. Bohn J, Eddings W, Schneeweiss S. Conducting privacy-preserving multivariable propensity score analysis when patient covariate information is stored in separate locations. Am J Epidemiol. 2017;185(6):501-510.
84. Christen P. Privacy-Preserving Record Linkage.; 2016. https://www.newton.ac.uk/event/dla. Accessed May 13, 2020.
85. Randall SM, Ferrante AM, Boyd JH, Brown AP, Semmens JB. Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581? HIM J. 2016;45(2):71-79.
86. Fernandes M, Decouchant J, Volp M, Couto FM, Esteves-Verissimo P. DNA-SeAl: Sensitivity Levels to Optimize the Performance of Privacy-Preserving DNA Alignment. IEEE J Biomed Heal Informatics. 2020;24(3):907-915.
87. Kuzu M, Kantarcioglu M, Inan A, Bertino E, Durham E, Malin B. Efficient Privacy-Aware Record Integration. Adv database Technol proceedings Int Conf Extending Database Technol. 2013:167-178.
88. Georgian. Georgian Impact Blog: A Brief Introduction to Differential Privacy. https://medium.com/georgian-impact-blog/a-brief-introduction-to-differential-privacy-eacf8722283b. Published August 31, 2018. Accessed April 12, 2021.
89. Bohn J, Eddings W, Schneeweiss S. Conducting privacy-preserving multivariable propensity score analysis when patient covariate information is stored in separate locations. Am J Epidemiol. 2017;185(6):501-510.
90. Boyens C, Krishnan R, Padman R. Privacy-Preserving Data Releases for Health Report Generation. Stud Health Technol Inform. 2004;107(0):1268-1272.
91. Gleim LC, Karim MR, Zimmermann L, et al. Enabling ad-hoc reuse of private data repositories through schema extraction. J Biomed Semantics. 2020;11(1):6.
92. Randall S, Brown A, Ferrante A, Boyd J. Privacy preserving linkage using multiple match-keys. Int J Popul Data Sci. 2019;4(15).
93. Chen F, Jiang X, Wang S, et al. Perfectly Secure and Efficient Two-Party Electronic-Health-Record Linkage. IEEE Internet Comput. 2018;22(2):32-41.
94. Bian J, Loiacono A, Sura A, et al. Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. JAMIA Open. 2019;2(4):562-569.

95. Malin B, Airoldi E. Confidentiality preserving audits of electronic medical record access. Medinfo. 2007;12(Pt 1): 320-324.

96. Yigzaw KY, Michalas A, Bellika JG. Secure and scalable deduplication of horizontally partitioned health data for privacy-preserving distributed statistical computation. BMC Med Inform Decis Mak. 2017;17(1):1.

97. Neubauer T, Heurix J. A methodology for the pseudonymization of medical data. Int J Med Inform. 2011;80(3):190-204.

98. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak. 2009;9(1):1-11.

99. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. J Biomed Inform. 2014;50:205-212.

100. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak. 2009;9(1).

101. Schnell R, Borgs C. Protecting Record Linkage Identifiers Using a Language Model for Patient Names. Stud Health Technol Inform. 2018;253:91-95.

102. Schnell R, Borgs C. Protecting Record Linkage Identifiers Using a Language Model for Patient Names. Stud Health Technol Inform. 2018;253:91-95.

103. Brown AP, Borgs C, Randall SM, Schnell R. Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets. BMC Med Inform Decis Mak. 2017;17(1):83.

104. Brown AP, Randall SM, Boyd JH, Ferrante AM. Evaluation of approximate comparison methods on Bloom filters for probabilistic linkage. Int J Popul data Sci. 2019;4(1):1095.

105. Durham E, Xue Y, Kantarcioglu M, Malin B. Private medical record linkage with approximate matching. AMIA Annu Symp Proc. 2010;2010:182-186.

106. Demelius L, Kreiner K, Hayn D, Nitzlnader M, Schreier G. Encoding of Numerical Data for Privacy-Preserving Record Linkage. Stud Health Technol Inform. 2020;271:23-30.

107. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials. 2015;16(1):1-14.

108. Zumel N. A Simpler Explanation of Differential Privacy. https://www.r-bloggers.com/2015/10/a-simpler-explanation-of-differential-privacy/. Published 2015.

109. Rubinstein BIP, Aldà F. diffpriv: An R Package for Easy Differential Privacy. Jounal Mach Learn Res. 2017;18:1-5. https://cran.r-project.org/web/packages/diffpriv/vignettes/diffpriv.pdf.

110. Loukides G, Gkoulalas-Divanis A. Utility-aware anonymization of diagnosis codes. IEEE J Biomed Heal informatics. 2013;17(1):60-70.

111. Guesdon M, Benzenine E, Gadouche K, Quantin C. Securizing data linkage in french public statistics. BMC Med Inform Decis Mak. 2016;16(1):129.

112. Durham E, Xue Y, Kantarcioglu M, Malin B. Quantifying the Correctness, Computational Complexity, and Security of Privacy-Preserving String Comparators for Record Linkage. Inf fusion. 2012;13(4):245-259.

113. Grobbee DiE, Hoes AW. Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research. Sudbury, Mass: Jones and Bartlett Publishers; 2014.

114. Airoldi EM, Bai X, Malin BA. An entropy approach to disclosure risk assessment: Lessons from real applications and simulated domains. Decis Support Syst. 2011;51(1):10-20.

115. Wagner I, Eckhoff D. Technical Privacy Metrics: A Systematic Survey. ACM Comput Surv. 2018;51(3).

116. Wang K, Phillips CA, Saxton AM, Langston MA. EntropyExplorer: An R package for computing and comparing differential Shannon entropy, differential coefficient of variation and differential expression. BMC Res Notes. 2015;8(1):832.

117. Baker D, Knoppers BM, Phillips M, et al. Privacy-Preserving Linkage of Genomic and Clinical Data Sets. IEEE/ACM Trans Comput Biol Bioinforma. July 2018.

118. Wang S, Bonomi L, Dai W, et al. Big Data Privacy in Biomedical Research. IEEE Trans big data. 2020;6(2):296-308.

119. Dong X, Randolph DA, Rajanna SK. Enabling Privacy Preserving Record Linkage Systems Using Asymmetric Key Cryptography. AMIA . Annu Symp proceedings AMIA Symp. 2019;2019:380-388.

120. Randall SM, Brown AP, Ferrante AM, Boyd JH. Privacy preserving linkage using multiple match-keys. Int J Popul data Sci. 2019;4(1):1094.

121. US Census Bureau. 2020 Census Operational Plan: A New Design for the 21st Century.; 2018.

122. Earle CC. Ontario Cancer Data Linkage Project: "cd-link." Healthc Q. 2014;17(1):11-13.

123. Brown AP, Randall SM. Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model. JMIR Med Informatics. 2020;8:1-15.

124. Pal D, Chen T, Zhong S, Khethavath P. Designing an algorithm to preserve privacy for medical record linkage with error-prone data. J Med Internet Res. 2014;16(1):1-17.

125. Stammler S, Kussel T, Schoppmann P, et al. Mainzelliste SecureEpiLinker (MainSEL): Privacy-Preserving Record Linkage using Secure Multi-Party Computation. Bioinformatics. 2020;(May):6-11.

126. Nasseh D, Stausberg J. Impact of variations in anonymous record linkage on weight distribution and classification. Stud Health Technol Inform. 2013;192:922.

127. Nguyen L, Stoové M, Boyle D, et al. Privacy-preserving record linkage of deidentified records within a public health surveillance system: Evaluation study. J Med Internet Res. 2020;22(6).

128. Agiro A, Chen X, Eshete B, et al. Data linkages between patient-powered research networks and health plans: A foundation for collaborative research. J Am Med Informatics Assoc. 2019;26(7):594-602.

129. Nasseh D, Engel J, Mansmann U, Tretter W, Stausberg J. Matching Study to Registry data: Maintaining Data Privacy in a Study on Family based Colorectal Cancer. Stud Health Technol Inform. 2014;205:808-812.

130. Wang S, Jiang X, Singh S, et al. Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. Ann N Y Acad Sci. 2017;1387(1):73-83.

131. Kho AN, Cashy JP, Jackson KL, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J Am Med Informatics Assoc. 2015;22(5):1072-1080.

132. Yu F, Fienberg SE, Slavković AB, Uhler C. Scalable privacy-preserving data sharing methodology for genome-wide association studies. J Biomed Inform. 2014;50:133-141.

133. Canterberry M, Kaul AF, Goel S, et al. The Patient-Centered Outcomes Research Network Antibiotics and Childhood Growth Study: Implementing Patient Data Linkage. Popul Health Manag. 2019;00(00):1-7.

134. Jones KH, Ford D V, Jones C, et al. A case study of the secure anonymous information linkage (SAIL) gateway: A privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform. 2014;50:196-204.

135. Schmidlin K, Clough-Gorr KM, Spoerri A. Privacy preserving probabilistic record linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. BMC Med Res Methodol. 2015;15:46.

136. Brown AP, Ferrante AM, Randall SM, Boyd JH, Semmens JB. Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage. Front public Heal. 2017;5:34.

137. Nasseh D, Stausberg J. Evaluation of a Binary Semi-supervised Classification Technique for Probabilistic Record Linkage. Methods Inf Med. 2016;55(2):136-143.

138. Stausberg J, Waldenburger A, Borgs C, Schnell R. Combining Different Privacy-Preserving Record Linkage Methods for Hospital Admission Data. Stud Health Technol Inform. 2017;235:161-165.

139. Loukides G, Gkoulalas-Divanis A, Malin B. Privacy-Preserving Publication of Diagnosis Codes for Effective Biomedical Analysis. ITAB 2010 Corfu, Greece 10th Int Conf Inf Technol Appl Biomed Emerg Technol Patient Specif Healthc 2-5 Novemb 2010, Aquis Corfu Holiday Palace Hotel Greece Int. 2010;2010:1-6.

140. Laud P, Pankova A. Privacy-preserving record linkage in large databases using secure multiparty computation. BMC Med Genomics. 2018;11.

141. Lu W-J, Yamada Y, Sakuma J. Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption. BMC Med Inform Decis Mak. 2015;15:S1.

142. Chen F, Jiang X, Wang S, et al. Perfectly Secure and Efficient Two-Party Electronic-Health- Record Linkage. IEEE Internet Comput. 2018;22(2):32-41.

143. Brown A, Randall S, Boyd J, Ferrante A. Evaluation of approximate comparison methods on Bloom filters for probabilistic linkag. Int J Popul Data Sci. 2019;4(16).

144. Cassa CA, Miller RA, Mandl KD. A novel, privacy-preserving cryptographic approach for sharing sequencing data. J Am Med Informatics Assoc. 2013;20(1):69-76.

145. Wang S, Bonomi L, Dai W, et al. Big Data Privacy in Biomedical Research. IEEE Trans Big Data. 2020;6(2):296-308.

SUPPLEMENTARY MATERIALS

Systematic review substantiating this PPRL guide

Methods

A query to search PubMed, Embase and Cochrane Library was developed by one author (WD). Adding “linkage” to the search query was considered, yet we opine that this limited the search results too much. The searches were not limited by time constraints. Databases were searched on latest on December 14, 2020.

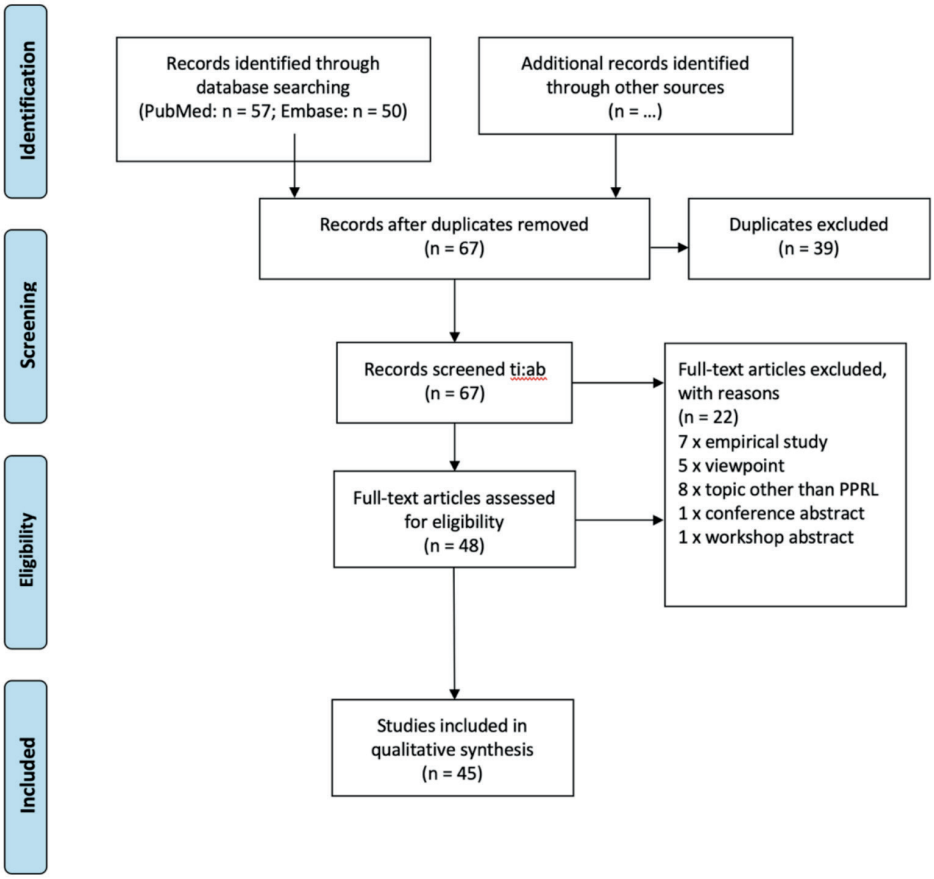
Executed query: “privacy”[tiab] AND (“preserving”[tiab] OR “protect*[tiab])

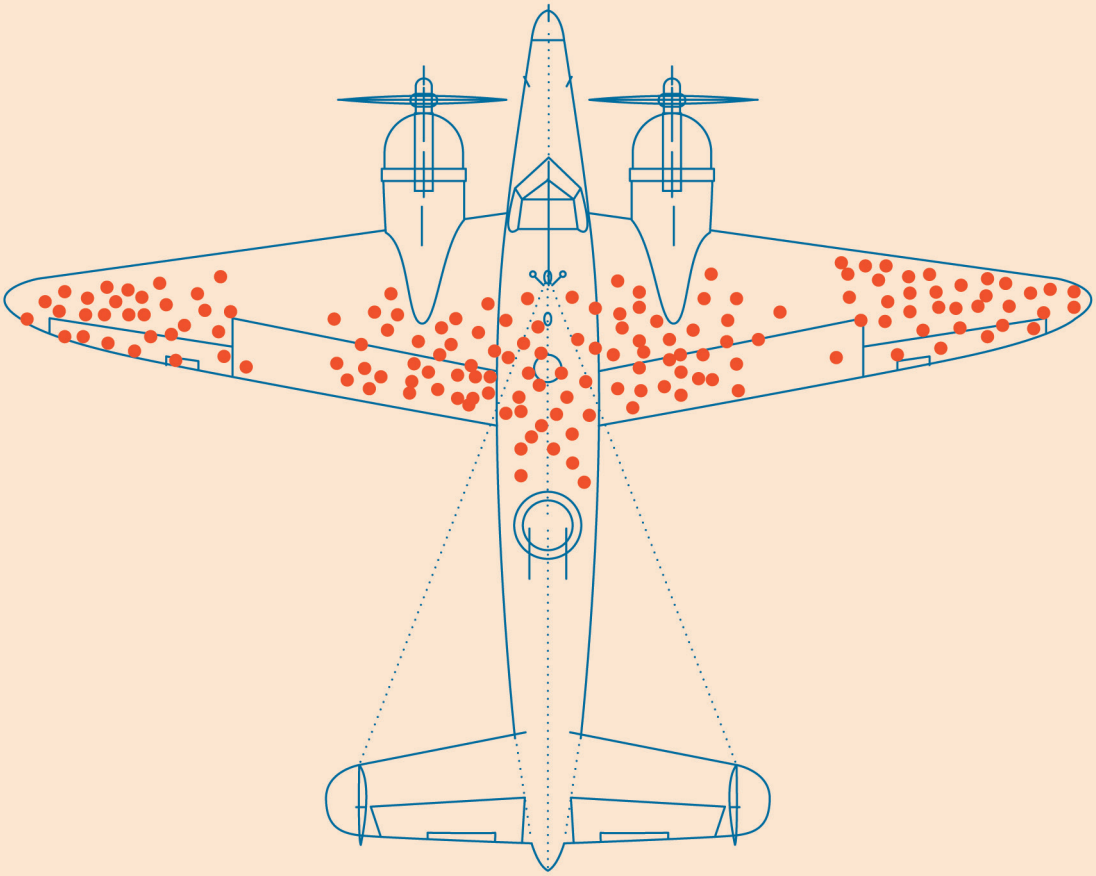
Results were imported in rayyan.ai. Titles, abstracts and full texts were screened by one author (WD). Articles describing (1) only record linkage, (2) only privacy-preserving and (3) combined PPRL methods, (4) performance of individual methods and (5) proofs of concept were included. Uncertainties were discussed with three other authors (ES, IV, MZ) to reach agreement on whether to include of exclude articles.

From included literature we extracted and categorized common and repeating themes. Articles generally considered records linkage, privacy preservation or both. As sufficient guidance on record linkage is available, we focused on articles that comprised both record linkage and privacy protection. Extracted themes were included in the guide in order of occurrence in the data linkage process based on literature and experiences of the authors.

In total, 1376 articles were found across PubMed, Embase and the Cochrane Library (Figure S1). Of these, 1107 articles remained after deduplication. Title, abstract and full text screening resulted in 45 articles to be included eventually. 2,3,22–24,26,27,29 31,33,34,39,4,40,50–58,9,59–68,11,69–73,12,14,15,18,20 Most articles were included because they did not describe record linkage, privacy preservation, nor a combination of these (N=XX).

Figure S1: Flow chart systematic literature search





Chapter 42

General discussion, future perspectives & valorization

In this general discussion, I will first revisit the main findings of the three main parts of this thesis. Next, I will discuss these findings along three coinciding viewpoints. In these viewpoints I will subsequently (1) put knowledge generation in a general perspective (2) discuss the context in which I think learning healthcare systems (LHSs) can be applied relevantly, and (3) address the use of routinely collected healthcare data (RCHD) to supplement conventional clinical trials and in the development and validation of prediction models. I will end this general discussion by outlining future perspectives for LHSs and discussing the valorization of this thesis.

In part 1, I covered the current state of evidence that serves as the basis for cardiovascular clinical practice. Evidence substantiating clinical knowledge is ranked (also known as the evidence hierarchy), where large randomized controlled trials (RCTs) and meta-analysis represent the highest (i.e., best) level (A), small RCTs and observational studies an average level (B), and case series and expert opinions the lowest (i.e. worst) level (C). When judged on these levels of evidence (LoE), the current cardiovascular knowledge base is thin (Chapter 2) and only applicable limitedly worldwide (Chapter 3).

In part 2, I covered the current state of LHSs in literature, the consequences that LHSs might have for the robustness and reliability of knowledge, and the duties of professionals in LHSs. A uniform interpretation of LHSs is missing in literature (Chapter 4). As a result, authors use the term LHS in many different contexts and is often cited as ann opportunistic term. Frequently, a sense of the larger concept of LHSs is lacking. Authors generally only describe a feature which they then state to be essential for LHSs. Such a wider sense of the concept of LHSs is important as LHSs might bring changes to the value of knowledge generated in LHSs (Chapter 5). LHSs might generate evidence faster and more tailored to their local system increasing local relevance of evidence. Yet, LHSs can also make evidence generated by them less relevant outside their system due to too much tailoring of evidence. Evidence generated for local use might also be subjected less to independent external peer review, making the evidence less reliable and more prone for errors. For professionals, LHSs might pose a dilemma. In LHSs the system is expected to be aligned around common value adding incentives. Consequently, it might be harder or less accepted for professionals to dissent from the systems choices. In particular in acute and subacute situations such as what the COVID-19 pandemic brought dissenting opinions might not always be valued and might be surmised to be undesirable (Chapter 6).

In part 3, I covered the use of RCHD as a promising method for LHSs. Text-mining in electronic healthcare records (EHRs) can be very valuable for LHSs. I found that text-mining in EHRs was suitable for participant screening, collecting baseline (Chapter 7) and outcome data (Chapter 8). As such, text-mining in EHRs can facilitate making research more affordable and scalable. Next to the usage of RCHD in trials they are increasingly used in prediction models to support clinical decision making. I researched the effect on performance when using prediction models in practice over a longer period. Over time, the EuroSCORE II prediction model proved to consistently overestimate mortality for cardiac surgery in the Netherlands when using data from the RCHD of the Netherlands Heart Registry (NHR). This systematic overestimation implies that it is important to review prediction models regularly, because they influence treatment decisions in clinical practice (Chapter 9). Lastly, I researched how the value of routine care datasets can be increased by linking them in a privacy preserving manner (Chapter 10). In the final chapter I provided a stepped plan to follow to link RCHD while maintaining subject anonymity.

1. SEPARATING EVIDENCE FROM KNOWLEDGE, THE RIGHT RESEARCH DESIGNS FOR THE RIGHT CONTEXTS

Corresponding to the findings of the first part of this thesis, evidence levels for many medical specialties have been found to be low for clinical practice.[1–6] While low levels of evidence seem disconcerting they do not always mean that our level of knowledge is low. Evidence levels came along with the introduction of evidence-based medicine (EBM), and should indicate the trust that can be given to recommendations derived from this evidence.[7] However, not all our knowledge is testable to the highest degree of evidence we currently adhere to, i.e. meta-analyses and RCTs.[8]

The authors of the editorial that accompanied the publication of the manuscript underlying the second chapter of this thesis rightly criticized that it is unlikely that "there will ever be RCTs confirming that a routine blood sample should be performed as soon as possible in the suspicion of acute coronary syndrome (LoE C), or that *parachutes can reduce the risk of injury"*.[8,9] As the previous example shows, evidence levels cannot be read as synonyms to the knowledge we have, but only indicate the strength of evidence IF available. Moreover, the example shows that generating higher levels of evidence to validate our knowledge can be grossly unethical. When 'strong' evidence is present the evidence strengthens the certainty of our knowledge.[10] However, when 'strong' evidence is absent this does not always indicate uncertainty of our knowledge, but may simply be explained by its

context and the fact that RCTs are not possible or ethical.[11] In diagnostic studies, for example, RCTs are generally not the appropriate method to research their accuracy.[12] To deal with this relativity of evidence, researchers and clinicians should appreciate evidence in its context. Dogmatic application of evidence levels can wrongly disqualify existing knowledge as non-existent and incorrectly determine high quality (observational) studies to be futile.[13]

It is true that RCTs remain the pinnacle of theoretical perfection of study designs and often are the best way to fill gaps in our knowledge on the benefits and risks of interventions, and thus to be used as reference standard.[14] It is hard, if not impossible, to methodologically challenge the randomization of patient characteristics used in RCTs to allow unbiased investigation of one determinant of interest.[14] Nonetheless, RCTs are often found not to be perfect either.[15] RCTs frequently report results that are different from those experienced in the 'real-world', i.e., clinical practice.[16] This difference comes from many different factors, among others that in clinical practice it is not possible to subscribe interventions only to patients who fulfil the eligibility criteria of trials .[17] Clinicians aim to help their patients, and will subscribe interventions to patients for who they think the intervention might help, which means they may apply the intervention to a broad population than originally studied.[18] In addition, therapeutic adherence of patients is generally found to be lower outside trial contexts.[17] Among others, the ethical limitations mentioned in the previous paragraph and research-practice mismatch make that conducting trials is not always the best way to answer questions longing for answers.

Observational studies can often be used besides RCTs to fill gaps in our knowledge.[19] Moreover, increased awareness of confounding in research makes well conducted observational studies more reliable than ever before.[20] Today, a wide range of methods exist to deal with result distortions that can be found in observational studies. When applied correctly, these methods can provide a high degree of protection against biases in study results too.[19,21] When no other large confounders can be thought of to be compensated for through propensity scores, mendelian randomizations or stratifications it is most likely that the most important confounders were caught.[22] Indeed, it might occur that missed large confounders are found retrospectively. Sound peer review of observational research before publication is therefore of even of greater importance than for randomized research. Fear for residual confounders, however, should not limit the use of observational studies in the large.

This is no dissertation promoting observational research over clinical trials, only a manuscript to emphasize the importance of putting evidence, knowledge, and research in perspective of its context. When we want to move toward LHSs many see the use of observational data as a prerequisite. Fear of using these data and skepticism for the studies conducted with observational data should therefore be reduced. For as long as we do not want to push people out of planes without parachutes some questions will require RCTs, while others can only be answered with common sense and observational data.[23] For diagnostic and prognostic research, observational studies are even the reference standard.[12] When evidence is missing it does not always imply knowledge is missing too. For clinical practice guideline committees this means that they must decide consciously whether they want to substantiate their recommendations to include best (high-quality) evidence only or best *available* evidence. Additionally, guideline committees should be aware that applicability of their recommendations depend on the context they are to be applied in, thus excluding the use of some recommendations in certain contexts when choosing for best available methods and treatments only.

2. LEARNING HEALTHCARE SYSTEMS ARE NOT THE ANSWER TO THE ULTIMATE QUESTION ~~OF LIFE, THE UNIVERSE AND EVERYTHING~~

> Mice are not what they appear to be. They are a group of hyper-intelligent pan-dimensional beings demand to learn the Answer to the Ultimate Question of Life, The Universe, and Everything.[24] To answer this question they built the supercomputer Deep Thought specifically for this. The answer? 42.

42

The questions for LHSs to answer

In literature, LHSs are often suggested as the answer to all our clinical, research and methodological challenges.[25] However, when questions are asked too broad, answers will also be too broad to be meaningful.

Indeed, it is probable that much more can be learned in a system of continuous improvement that leverages clinical data to learn from.[26] For example, RCHD including EHR data, registry and insurance data is estimated to be used for outcome data collection in less than 1.5% of RCTs undertaken still.[27] As results from this thesis show, RCHD

can likely be used more often in trials. This means, with the proper data infrastructure, LHSs will be able to facilitate trial conduct, thereby resulting in trials that cannot only be conducted cheaper, but also faster.[28] Another example is the use of patient-reported outcome measures (PROMs; albeit often not present in RCHD).[29] Despite their usage being on the rise, use of PROMs in research is still limited.[30] Because of the close patient-clinician relationship advocated in LHS, and considering that clinicians are often the primary initiators of research, LHSs will be able further strengthen the input of patients in clinical research.

Besides advantages, LHSs will have disadvantages also. Speeding up implementation of research will leave less time for its results to be validated by the passing of time before being implemented into care (see Chapter 5). Some research conducted in LHSs might also have a narrower scope, limiting their relevance and applicability to the LHSs itself instead of a broader clinical context. Moreover, new methods that allow instant data analysis from very large amounts of data are very much in development themselves still.[31] In particular, how and when we can make causal inferences from (observational) RCHD is much debated still.[23,26] These effects of LHSs on evidence and knowledge generation can often be viewed as a difference between research and science. According to the Oxford English Dictionary research is the *"careful study of a subject",* while science considers "*knowledge of the natural and physical world*", which could be operationalized in that research is more focused on addressing local needs, while science focusses on more universal, larger questions.[32] Depending on the way LHSs are brought into practice, LHSs do not have to replace EBM, but can serve as an evidence generating option to facilitate and speed up EBM.

Attention for these disadvantages of LHSs is still required, as to easily ignoring them can bear risks for clinical practice when internally focused research (e.g., on locally applied clinical prediction models) is validated less by external reviewers. To guard against these risks validation of research results by independent reviewers and placing the results into the broader body of evidence will remain of utmost importance.

Focus on processes next to methods to allow answering the right questions

Many of the challenges for LHSs are not related to methods of research or care but rather to their governing processes. The idea of integrating research with care in LHSs is built on the premise that research and care function as systems. A system inherently assumes a top-down perspective. Research is grounded in a system of accumulating evidence

that allows us to distil valuable knowledge from flawed assumptions.[33] Care allows us to help patients by use of our system of payers and providers. Patients can rely on insurances and visit hospitals. Clinicians in hospitals can order all sorts of additional tests and interventions to help care for their patients. However, for LHSs a structured system approach appears to be still missing. Individual researchers cite LHSs when convenient, often without regard for the larger system (see Chapter 4).

To set a systems approach for LHSs into motion, the four characteristics of LHSs should be regarded as requirements instead (see Chapter 4 for details on these characteristics and how they are used in literature today). Regarding the characteristics as requirements will force leadership buy-in in systems to be changed into LHSs. First, incentives should change accordingly by aligning financial incentives with LHS aims, rewarding cooperative achievements of care-research combinations beneficial for the system instead of the individual or just a single pillar. Second, supporting infrastructures should be installed to allow science and informatics to use RCHD and support real-time learning. Third, clinicians and their patients should be trained in shared decision making to increase relevancy of outcomes researched by the system. Fourth, a continuous learning culture should be cultivated by changing *fixed mindsets* to *growth mindsets*.[34] To attain these requirements leadership of healthcare systems will have to put words into actions top-down. This does not mean that healthcare leaders should dictate research topics in LHSs, but it does mean that they provide the prerequisites of LHSs and that they reward research on topics more relevant for an LHS.

The levels of LHSs to ease finding the right questions

To make it easier to determine which questions should be answered in LHSs clearer boundaries of where 'systems' start, and end are needed. This void makes that LHSs are at risk of comparing apples with oranges. For example, aims of nationwide LHSs are likely to be different from aims of departmental LHSs. In nationwide LHSs aims are likely to be focused on improving population outcomes with adequate incentives. In departmental LHSs, on the other hand, aims will be more focused on personalized outcomes for patients with adequate information technology infrastructures.

LHSs should be distinguished from each other based on their aims and level in society. Currently LHSs are not separated in such levels. As a result, it is hard for researchers and clinicians to find literature relevant for their context. And more importantly, it is hard for all parties involved to know on which topics more research is needed still. Separating LHSs

into different levels also protects them from being seen as the answer to all questions, as it allows involved parties to see the forest for the trees again.

3. OPPORTUNITIES FOR USING ROUTINELY COLLECTED ELECTRONIC HEALTHCARE RECORD DATA IN CLINICAL TRIALS

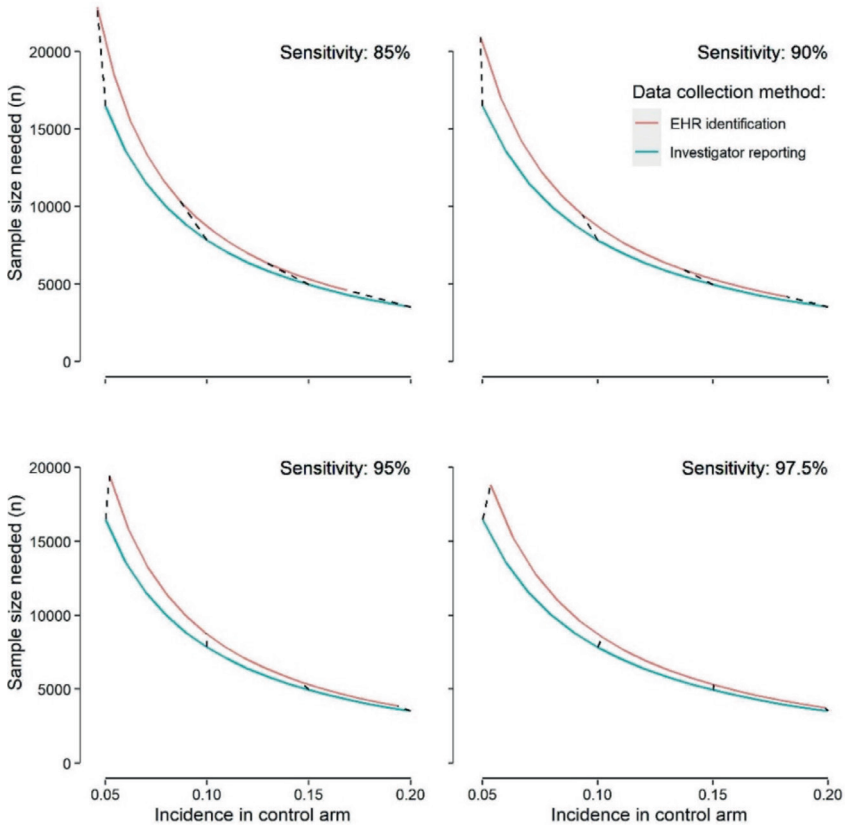
When, about a decade ago, EHRs were introduced they were meant to ease administration processes of healthcare professionals.[35] By entering files into a computer it was easier to find and collect patient' health records than when doing this manually. Moreover, EHRs could make the reimbursement of expenses less laborious. Today, we want to do much more with EHRs than just being a digital form of paper. We want EHRs to be able to follow the medical journey of patients and alert us when patients in our care are at risk for some adverse outcome.[36–38] In this manner EHRs can serve as an ideal basis for data collection in LHSs. For example, I found that EHRs can be used to screen patients on trial eligibility, lowering the number of patients to be screened for trial participation by almost 80%, having the potential to accelerate participant recruitment and lower associated costs substantially (Chapter 7).[40] Also, I showed that we could identify first and secondary endpoints of a recent cardiovascular trial, the LoDoCo2, with an average sensitivity and specificity of 87% and 85% respectively, having the potential to lower the laboriousness of data collection in clinical trials (Chapter 8). Sensitivity and specificity were higher when endpoints were more explicit.

However, using EHR data in a randomized trial comes at a price. RCHD is less precise than current standardized measurements, leading to measurement error of trial endpoints.[41] What is more, (binary) endpoints may be missed, thus leading to misclassification of endpoints.[42,43] Even if measurement error or misclassification is independent of treatment status, it will almost always reduce the power of a trial, and if dependent on treatment status may introduce a bias of the treatment effect estimate.[43]

When the power of a trial is reduced, more participants are needed to detect statistically significant differences across study arms. The two most important aspects when calculating the required sample size of a trial are the incidence of the endpoint in the overall population under standard care, which depends on the sensitivity to detect a given endpoint, and the relative difference in incidence between the study arms (assuming standard practice of an alpha of 0.05 and beta of 0.20). The usage of EHR data in a clinical trial could in effect lower the sensitivity of endpoint detection. In **Figure 1**, four examples

are given of effects of various endpoint sensitivities on the power of a trial. As endpoint sensitivity differs, the usage for EHR data for detecting them thus may or may not be advantageous.

Figure 1: Effects on statistical power of trials with relative differences of sensitivities



When and in which contexts can EHR data play a role for clinical trials?

For participant recruitment, EHR data can already be used for all trials. Screening patients automatically will decrease efforts needed to recruit participants as it can be done in a more targeted fashion. If patients are selected that are part of the study domain only benefits can be expected from automating (parts) of the recruitment process.

For data collection during follow-up, it depends whether EHR data are suitable to be used. Trials that are suitable for using EHR data should preferably focus on (pragmatic) effectiveness rather than efficacy. In general, efficacy studies are used for drug

development studies, placing most of these studies outside the realm of EHR research as they will not be conducted in clinical practice but in controlled environments. In addition, as the characteristics of new drugs will not be known, data collection should be as accurate as possible notably for safety outcomes. Trials that are suitable for using EHR data should focus on interventions for which the safety profile does not belong to the key study objects (e.g., statins, colchicine). As some endpoints might be missed when using EHR data it is important that safety endpoints are not of interest to the study since missing these might be of danger to study participants. And these trials should use endpoints that are well captured in EHR data (e.g., MACE, mortality). In essence, use of EHR data can lower the burden of trials for both the participants and the researchers.

For trials, a larger sample size may be needed (to preserve power) when using routinely collected endpoints. However, this may be worth it when the use of EHR data facilitates recruitment and the burden to participate is low. When per participant costs and expected savings are known a benefit-loss budget could be calculated, quantifying the costs expected to be saved per power percentage point.

FUTURE PERSPECTIVES: TOWARDS LEARNING HEALTHCARE SYSTEMS

To transform current healthcare systems into LHSs further steps need to be taken. Consciousness of continuous learning is on the rise in healthcare systems, yet most healthcare systems are no LHSs. The ill definition and wide usage of the LHS term are its greatest strength and weakness. It is a strength because it makes it easy to adopt the term in research and care, and a weakness because it can deflate the term to a hype.

Nearing the end of this thesis, I would like to highlight two major aspects of LHSs that, in my opinion, need attention soon to move current healthcare systems into LHSs.

Leadership connecting research and clinical practice through supportive systems

Although one of the chief aims of LHSs is to integrate research and care, it has proven to be challenging to integrate both. A challenge most visible in the limited number (or absence dependent on one's view) of LHSs to date.[25] Notably, I found very limited research on the role of leadership in LHSs. Yet, as researchers are trained to do research and clinicians to deliver patient care, I believe healthcare leaders are the appropriate people to serve as

accelerators for LHSs. Therefore, research into the role of healthcare leaders would be of great value to investigate this hypothesis.

The convergence of epidemiology, biostatistics, and data science

Epidemiologists receive training on study designs, measurement methods and causal inference. However, when using ever bigger data sets with more complex relations, epidemiological methods alone are not enough to answer the questions at hand.[44] To be able to fully leverage the possibilities of big data and computer processing power, epidemiology, biostatistics and data science will have to converge. For epidemiologists, biostatisticians bring rigor on theory and application of statistical techniques. Data scientists bring expertise on advanced computations and visualizations. Taken together these three fields comprise all knowledge required to harness the full capabilities new techniques as artificial intelligence will bring us. To date, these fields often still operate independently and largely unaware of each other. Yet, to solve the challenges ahead they need to converge. Research in how these three disciplines could and should converge would therefore be of value to LHSs.

VALORIZATION

Karl Marx defined valorization as when something proves its worth in earnings or a yield.[45] Therefore, I cannot valorize the research conducted for this thesis by definition. However, I can report on how I hope that this research will add value soon.

First, I hope this thesis will add to the ongoing debate on LHSs by opening conversations on how we define LHSs and how we should govern them. LHSs can probably exist in too many forms to imagine, yet conscious considerations in how to achieve the goals for which LHSs where conceptualized will remain necessary. In this it is important to factor into account that LHSs will not be the answer to everything. LHSs should be considered as means, not goals.

Second, I hope this thesis will stimulate the use of RCHD in research and practice. In my opinion the question is not whether we are going to use RCHD, but when. The two showcases of RCHD use in clinical trials show that it is possible to use RCHD for data collection and what caveats one can encounter. Although, I do realize that more research on the use of RCHD is needed to assess when its benefits outweigh its disadvantages, I believe it is already possible to use RCHD in some situation.

CONCLUSION

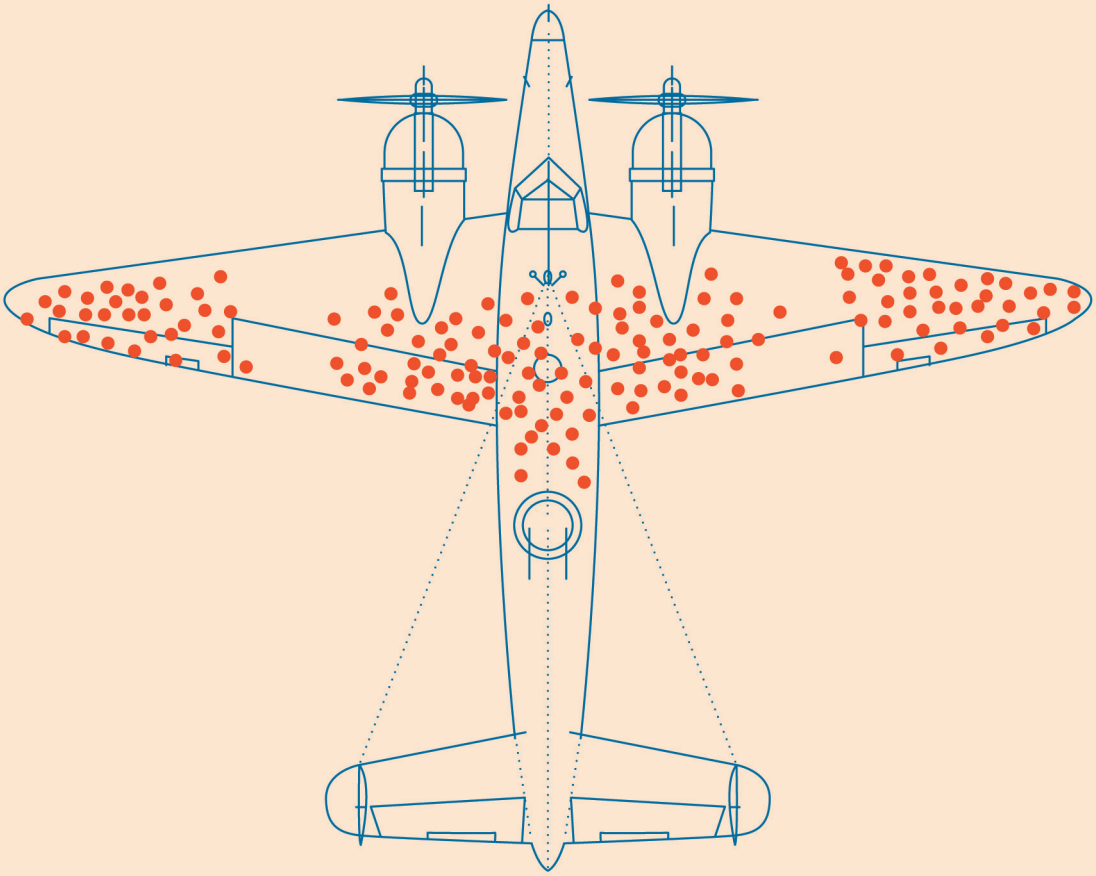
Methodologies used for knowledge generation in the context of LHSs will have consequences for the value of generated knowledge and the conduct of research.

In these concluding remarks, I want to highlight three aspects of these consequences. First, methodologies used for knowledge generation in LHSs will make research faster, more relevant in a local LHS, yet in some situations less valuable outside the LHS it was created in creating a difference in value inside and outside LHSs. Second, the use of RCHD for knowledge generation in LHSs will change how research is conducted. Lowering labor required for the conduct of research and increasing efficiency as a result. Third, LHSs will require frequent, independent monitoring of research findings to adequately ensure high quality use of knowledge.

REFERENCES

1. Alseiari M, Meyer KB, Wong JB. Evidence Underlying KDIGO (Kidney Disease: Improving Global Outcomes) Guideline Recommendations: A Systematic Review. Am J Kidney Dis. 2016;67(3):417-422.
2. Lee DH, Vielemeyer O. Analysis of Overall Level of Evidence Behind Infectious Diseases Society of America Practice Guidelines. Arch Intern Med. 2011;171(1):18-22.
3. Venkatesh AK, Savage D, Sandefur B, Bernard KR, Rothenberg C, Schuur JD. Systematic review of emergency medicine clinical practice guidelines: Implications for research and policy. Woolfall K, ed. PLoS One. 2017;12(6):e0178456.
4. Feuerstein JD, Akbari M, Gifford AE, et al. Systematic analysis underlying the quality of the scientific evidence and conflicts of interest in interventional medicine subspecialty guidelines. Mayo Clin Proc. 2014;89(1):16-24.
5. Meyer C, Bowers A, Wayant C, et al. Scientific evidence underlying the American College of Gastroenterology's clinical practice guidelines. Grimmer KA, ed. PLoS One. 2018;13(10):e0204720.
6. Poonacha TK, Go RS. Level of scientific evidence underlying recommendations arising from the National Comprehensive Cancer Network clinical practice guidelines. J Clin Oncol. 2011;29(2):186-191.
7. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet. 2017;390(10092):415-423.
8. Piepoli MF, Ceconi C. Levels of evidence in the European Society of Cardiology Guidelines: Gaps in knowledge? Eur J Prev Cardiol. August 2019:204748731986941.
9. van Dijk WB, Grobbee DE, de Vries MC, Groenwold RHH, van der Graaf R, Schuit E. A systematic breakdown of the levels of evidence supporting the European Society of Cardiology guidelines. Eur J Prev Cardiol. 2019;26(18):1944-1952.
10. Hultcrantz M, Rind D, Akl EA, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol. 2017;87:4-13.
11. Dizon JM, Machingaidze S, Grimmer K. To adopt, to adapt, or to contextualise? The big question in clinical practice guideline development. BMC Res Notes. 2016;9(1):442.
12. Biesheuvel CJ, Grobbee DE, Moons KGM. Distraction From Randomization in Diagnostic Research. Ann Epidemiol. 2006;16(7):540-544.
13. Institute of Medicine. Observational Studies in a Learning Health System. Washington, DC; The National Academies Press; 2013.
14. Grobbee DiE, Hoes AW. Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research. Sudbury, Mass: Jones and Bartlett Publishers; 2014.
15. Ferguson RE, Fiore L. The Changing Face of Clinical Trials - Pragmatic trials. N Engl J Med. 2016;375(5):454-463.
16. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials. 2015;16(1):1-14.
17. Panayidou K, Gsteiger S, Egger M, et al. GetReal in mathematical modelling: a review of studies predicting drug effectiveness in the real world. Res Synth Methods. 2016;7(3):264-277.
18. Averitt AJ, Weng C, Ryan P, Perotte A. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. npj Digit Med 2020 31. 2020;3(1):1-10.
19. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet. 2004;363(9422):1728-1731.
20. Lipsitch M, Jha A, Simonsen L. Observational studies and the difficult quest for causality: Lessons from vaccine effectiveness and impact studies. Int J Epidemiol. 2016;45(6):2060-2074.
21. Dahabreh IJ, Sheldrick RC, Paulus JK, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. Eur Heart J. 2012;33(15):1893-1901.

22. Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. Am J Epidemiol. 2007;166(6):646-655.
23. Imai K, King G, Stuart EA. Misunderstandings between experimantalists and observationalists about causal inference. J R Stat Soc. 2014; 171(2):481-502.
24. Adams D. The Hitchhiker's Guide to the Galaxy. In: Pocket Books; 1979.
25. Budrionis A, Bellika JG. The Learning Healthcare System: Where are we now? A systematic review. J Biomed Inform. 2016;64:87-92.
26. Dahabreh IJ, Kent DM, IJ D, DM K. Can the learning health care system be educated with observational data? JAMA. 2014;312(2):129-130.
27. Lensen S, Macnair A, Love SB, et al. Access to routinely collected health data for clinical trials - Review of successful data requests to UK registries. Trials. 2020;21(1):398.
28. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med. 2011;104(12):510-520.
29. Maddox TM, Albert NM, Borden WB, et al. The Learning Healthcare System and Cardiovascular Care: A Scientific Statement from the American Heart Association. Circulation. 2017;134(14):e826–e857.
30. Howie L, Hirsch B, Locklear T, Abernethy AP. Assessing The Value Of Patient-Generated Data To Comparative Effectiveness Research. Health Aff. 2014;33(7):1220-1228.
31. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: Are we there yet? Heart. 2018;104:1156-1164.
32. Oxford English Dictionary. In: Oxford: Oxford University Press; 2021.
33. Bauer H. Ethics in Science: The Knowledge Filter. https://www.tissuegroup.chem.vt.edu/chem-ed/ethics/hbauer/hbauer-filter.html. Published 1995. Accessed December 20, 2019.
34. Dweck C. Mindset: The New Psychology of Success. New York: Random House; 2006.
35. Atherton J. Development of the electronic health record. Virtual Mentor. 2011;13(3):186-189.
36. Cimino JJ. Improving the Electronic Health Record: Getting What We Wished For. JAMA. 2013;309(10):991-992.
37. van Staa TP, Dyson L, McCann G, et al. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: Evaluations of two exemplar trials. Health Technol Assess (Rockv). 2014;18(43).
38. Hemingway H, Asselbergs FW, Danesh J, et al. Big data from electronic health records for early and late translational cardiovascular research: Challenges and potential. Eur Heart J. 2018;39(16):1481-1495.
39. van Staa TP, Klungel O, Smeeth L. Use of electronic healthcare records in large-scale simple randomized trials at the point of care for the documentation of value-based medicine. J Intern Med. 2014;275(6):562-569.
40. van Dijk WB, Fiolet ATL, Schuit E, et al. Text-mining in electronic healthcare records can be used as efficient tool for screening and data-collection in cardiovascular trials: a multicenter validation study. J Clin Epidemiol. 2020;132(1):97-105.
41. Williams R. The Added Value of Linked Data in Pharmacoepidemiology: Solving the Jigsaw Puzzle; Utrecht: Utrecht University; 2018.
42. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol. 2012;65(3):343-349.e2.
43. Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. J Clin Epidemiol. 2018;98:89-97.
44. Goldstein ND, Levasseur M, Mcclure LA. On the Convergence of Epidemiology, Biostatistics, and Data Science. Harvard Data Sci Rev. 2020;2(2).
45. Marx K. Das Kapital. (Engels F, ed.). Washington, DC: Regnery Publishing; 1996.



Addenda

Summary
Nederlandse samenvatting
List of publications
Dankwoord
About the author

SUMMARY

Regularly, meta-research finds that medical research is disconnected from the clinical practice it seeks to assist. Inversely, it is hard for researchers to acquire data from clinical practice to learn from. The disconnect between research and clinical practice results in the science-to-care gap that inspired the research for this thesis on learning healthcare systems (LHSs). In particular, the use of (new) methodologies to (re)connect research and care might also have consequences for the way knowledge is generated. In this thesis we aimed to investigate the consequences of using LHS methodologies for knowledge generation in the healthcare context.

In the first part of this thesis, we investigated the current state of knowledge in cardiovascular care. In **Chapter 2** we showed that evidence levels substantiating clinical practice guideline recommendations on cardiovascular care are still moderately low. Of the strong recommendations[1] issued by the European Society of Cardiology (ESC), less than fifteen percent of the recommendations is supported by the highest level of evidence. Recommendations on therapeutic topics, in particular on pharmaceutical and minimal invasive procedures, comprised more recommendations supported by high level evidence than recommendations on diagnostic topics. Moreover, a substantial number of recommendations on topics unrelated to clinical practice were found (e.g. recommendations on political topics).

The global availability and applicability of ESC guideline recommendations was investigated in **Chapter 3**. We found that actions recommended in ESC guidelines were more available in high income countries compared to lower and low-income countries. The availability of recommended actions and applicability of guideline recommendations in countries was negatively correlated with the gross national income of countries. Noticeable, some recommended actions emerged as limited in availability regardless country income levels. When asked for reasons of limited availability of recommend actions financial reasons for healthcare systems, institutions and patients were named most often.

In the second part of this thesis, we investigated the current state of implementation of LHSs. LHSs were conceptualized in 2007 already. In **Chapter 4** we investigated the current

1 Strong recommendation in this context comprises recommendations that positively and negatively recommend the use of treatments and procedures, opposed to recommendations that recommend consideration of treatments and procedures.

literature base on LHSs to assess where LHSs are today on methodological and ethical topics. We found that literature on LHSs is scattered and follows the key characteristics as outlined by the Institute of Medicine limitedly. Generally, articles do not go into details of methods or ethical concepts. The concept of LHSs often appeared to be used as a general desired future state without awareness of how this future state should or would look like in practice.

In **Chapter 5** we investigated how the social value of knowledge generated in LHSs might change compared to knowledge that is currently generated. We found that social value can be broken down in internal, external, and temporal social value. The internal value of knowledge (the value of knowledge for the system the knowledge is generated in) will likely increase in LHSs, as research performed in LHSs can be tailored to local system needs and can be generated and implemented faster. This might also increase patient engagement in research in these systems, as patients will find the results more relevant for themselves. Contrarily, the external value of knowledge (the value of knowledge for the society at large) might decrease. When research focuses too much on internal interests (e.g. by creating very specialized prediction models), the generated knowledge might be less valuable for other systems than it was generated in. The temporal value of knowledge generated in LHSs in general might also decrease. As evidence is generated and applied faster, the system and society in general has less time to validate the research results that substantiate the knowledge. Rendering the chance of applying erroneous research results in clinical practice higher.

The COVID19 pandemic brought to light some interesting aspects of LHSs. In the pandemic, research results were published very fast which inspired numerous debates on the validity of generated knowledge. As result, a Duty to Support LHSs was proposed in literature, stating that *experts should have the duty to support, and not undermine, health systems*. In response to this duty, we argued in **Chapter 6** that instead of putting constraints on the scientific debate, health systems should rearrange themselves to learn from opposing opinions to improve the way they conduct and communicate their research.

In the third part of this thesis, we investigated the usage of routinely collected healthcare data (RCHD) in LHSs. The use of RCHD in clinical trials can potentially make trials more affordable and less laborious. In **Chapter 7** we investigated how the use of RCHD from electronic healthcare records (EHRs) might increase efficiency of trial participant recruitment and baseline data collection. We found that automated screening of patients

for trial eligibility can reduce manual screening efforts by almost three quarters. Moreover, we found that baseline data can be collected with reasonable accuracy from structured and unstructured EHR data with automated data collection methods.

In **Chapter 8** we found that endpoint data collection of clinical cardiovascular trials can also benefit from automated data collection. For unambiguous cardiovascular endpoints (e.g. MACE) we were able to collect endpoints with a sensitivities of over 90% and a specificities of over 85%. For more ambiguous endpoints (e.g. ischemia-drive coronary revascularization) we found substantially lower sensitivities (just below 75%). Also, time series comparison of automatically collected endpoints and investigator reported endpoints showed high comparability, again more for unambiguous than for ambiguous endpoints.

Next to the usage of RCHD in trials, RCHD are increasingly used in prediction models to support clinical decision making as well. In **Chapter 9** we found that predictions models, in our study the EuroSCORE II, drift from their original calibration with the passing of time when not updated. The EuroSCORE II model proved to consistently overestimate in-hospital mortality after major cardiothoracic surgery in the Netherlands. Notable trends underlying this drift were a decrease in elective surgery, a rise in patients with lower NYHA heart failure classes and rise in patients with a recent myocardial infarction.

In the last chapter, **Chapter 10**, we developed a step-by-step guide to link multiple RCHD data sets in a privacy preserving manner. As RCHD are used more, it becomes more common to link multiple data sets. Preserving the privacy of subjects in data sets in linking efforts is increasingly important due to stricter regulations. To assist researchers in their privacy preserving linkage efforts we handed them a guideline outlining the major steps required in privacy-preserving record linkage.

NEDERLANDSE SAMENVATTING

Meta-onderzoek vindt regelmatig dat medisch onderzoek niet goed aansluit op de klinische praktijk die het onderzoek probeert te verbeteren. Voor onderzoekers is het daarentegen vaak lastig om gegevens uit de klinische praktijk te bemachtigen zodat zij van deze gegevens kunnen leren om de klinische praktijk te verbeteren. De matige aansluiting tussen onderzoek en klinische praktijk zorgt voor een onderzoeks-naar-zorg gat, het fenomeen wat het onderzoek voor deze dissertatie over lerende zorgsystemen (learning healthcare systems, verder gerefereerd aan als LHSs) inspireerde. De belangstelling in deze dissertatie ging in het bijzonder uit naar het gebruik van (nieuwe) methoden om onderzoek en zorg (weer) op elkaar aan te laten sluiten en de gevolgen die het gebruik van deze methoden kan hebben voor de waarde van kennis die het genereert. Deze dissertatie richtte zich op de gevolgen van het gebruik van LHS-methoden voor het genereren van kennis in de zorgcontext.

Voor het eerste deel van deze dissertatie onderzochten we de huidige staat van kennis over cardiovasculaire zorg. In **Hoofdstuk 2** toonden we aan dat bewijs niveaus (evidence levels) die ten grondslag liggen aan klinische praktijk richtlijn aanbevelingen over cardiovasculaire zorg nog steeds relatief laag zijn. Van de sterke aanbevelingen[2] die de European Society of Cardiology (ESC) uitvaardigt is minder dan vijftien procent onderbouwd met het hoogste bewijsniveau. Aanbevelingen over therapieën, in het bijzonder die over farmaceutische en minimaal-invasieve behandelingen, bevatte meer aanbevelingen ondersteund met hoge bewijsniveau's dan aanbevelingen over diagnostische onderwerpen. Daarnaast vonden we een substantieel aantal aanbevelingen die niet direct relevant waren voor de klinische praktijk. Dit betrof aanbevelingen over bijvoorbeeld politieke onderwerpen.

De wereldwijde beschikbaarheid en toepasbaarheid van ESC-richtlijn aanbevelingen onderzochten we in **Hoofdstuk 3**. We vonden dat acties aanbevolen in ESC-richtlijnen beter beschikbaar waren in landen met hoge inkomens dan in landen met lagere en lage inkomens. De beschikbaarheid van aanbevolen acties was negatief gecorreleerd met de inkomens van landen. Opmerkelijk was dat sommige aanbevelingen niet beschikbaar waren ongeacht het inkomensniveau van landen. Gevraagd naar redenen voor het niet of

2 Met sterke aanbevelingen wordt in dit geval aanbevelingen om behandelingen en procedures wel en niet te doen, deze staan hier tegenover aanbevelingen tot het in overweging nemen van behandelingen en procedures.

beperkt beschikbaar zijn van aanbevolen acties werden vooral financiële redenen voor het zorgsysteem, instituten en patiënten genoemd.

Voor het tweede deel van deze dissertatie onderzochten we de huidige staat van de implementatie van LHSs. Het concept LHS werd in al 2007 geintroduceerd, in **Hoofstuk 4** onderzochten we de huidige literatuur over LHSs om te kijken waar methoden en ethische aspecten van LHSs vandaag staan. We vonden dat de literatuur over LHSs versnipperd is en de sleutel karakteristieken zoals uiteengezet door het Amerikaanse Instiute of Medicine slechts beperkt volgden. Daarnaast leek het er vaak op dat het concept van LHSs gebruikt werd als ongedefinieerde toekomstige staat, zonder bewustzijn over hoe deze toekomst er concreet uit zou kunnen of moeten zien.

In **Hoofdstuk 5** onderzochten we hoe de sociale waarde van kennis gegenereerd in LHSs zou kunnen veranderen ten opzichte van vandaag de dag gegenereerde kennis. We vonden dat de sociale waarde van kennis uitgesplitst kon worden in de interne, externe en temporele sociale waarde van kennis. De interne waarde van kennis (de waarde die kennis heeft voor het systeem waar het in gegenereerd is) zal in LHSs waarschijnlijk toenemen voor de LHSs waar de kennis in gegenereerd is. Deze toename komt doordat de kennis op maat voor de wensen, uitvoering en implementatie van het systeem gegenereerd kan worden. Kennisgeneratie in LHSs zou tot meer patiënt betrokkenheid kunnen leiden doordat zij de relevantie van onderzoeksresultaten beter zouden kunnen zien. In tegenstelling tot de interne waarde van kennis, zou de externe waarde van kennis (de waarde die kennis heeft voor de bredere maatschappij) kunnen afnemen. Als onderzoek zich te veel op de interne interesses en belangen van LHSs richt kan dit resulteren in kennis die zo toegespitst zijn op het lokale LHS dat deze enkel in dat systeem toepasbaar is. De temporele waarde van kennis gegenereerd in LHSs zou ook af kunnen nemen. Als gevonden onderzoeksresultaten sneller toegepast worden in de praktijk blijft er minder tijd over om deze resultaten extern te valideren waardoor de kans dat foutieve resultaten toch geïmplementeerd worden toeneemt.

De COVID19 pandemie bracht een aantal interessante aspecten van LHSs aan het licht die tot daarvoor beperkt belicht waren. In de pandemie werden onderzoeksresultaten zeer snel gepubliceerd wat leidde tot verscheidene debatten over de validiteit van de gegenereerde kennis. Dit resulteerde in de literatuur tot het voorstel voor een Plicht LHSs te steunen (Duty to Support LHSs) waarin werd gesteld dat *experts de plicht zouden moeten hebben zorgsystemen te ondersteunen en deze niet mogen ondermijnen*. In reactie tot deze

plicht beargumenteerden wij in **Hoofdstuk 6** dat in plaats van het wetenschappelijk debat te limiteren zorgsystemen zichzelf zouden moeten herschikken om te leren van afwijkende meningen door onderzoeksuitvoering en communicatie te verbeteren.

Voor het derde deel van deze dissertatie onderzochten we het gebruik van routinematig verzamelde zorggegevens (routinely collected healthcare data, hier gerefereerd aan als RCHD) in LHSs. Het gebruik van RCHD zou klinische experimenten (trials) mogelijk betaalbaarder en minder arbeidsintensief kunnen maken. In **Hoofdstuk 7** onderzochten we hoe het gebruik van RCHD uit elektronische patiëntendossiers (EPD's) de efficiëntie van het werven van trial deelnemers en het verzamelen van baseline gegevens zou kunnen verhogen. We vonden dat het automatisch screenen van patiënten voor trial deelname geschiktheid, de benodigde hoeveelheid patiënten die handmatig gescreend moesten worden met bijna driekwart kon verminderen. Ook vonden we dat baseline gegevens automatisch verzameld kon worden vanuit gestructureerde en ongestructureerde EPD-gegevens met redelijk accuratesse.

In **Hoofdstuk 8** vonden we dat eindpunt gegevensverzameling voor klinische cardiovasculaire trials ook kan profiteren van geautomatiseerde gegevensverzameling. Voor scherp gedefinieerde eindpunten zoals MACE waren we in staat om eindpunten te verzamelen met sensitiviteit boven de 90% en specificiteit boven de 85%. Voor minder scherp gedefinieerde eindpunten zoals ischemie-gedreven coronaire vascularisatie vonden we substantieel lagere sensitiviteit (net onder de 70%). Tijdserie vergelijkingen lieten ook zien dat resultaten van geautomatiseerde gegevensverzameling resultaten zeer gelijkend op handmatige onderzoeker gerapporteerde eindpunten, wederom waren deze resultaten beter voor scherp gedefinieerde eindpunten dan voor minder scherp gedefinieerde eindpunten.

Naast het gebruik van RCHD voor trials worden RCHD in toenemende mate gebruikt in predictiemodellen die gebruikt worden beslissingen in de klinische praktijk te ondersteunen. In **Hoofdstuk 9** vonden we dat predictiemodellen, in ons geval de EuroSCORE II, over tijd gaan afwijken van hun uitgangswaarden. De EuroSCORE II liet zien consequent de ziekenhuismortaliteit in Nederland na cardiothoracale operaties te overschatten. Onderliggende trends die wij zagen in factoren van de EuroSCORE II waren een afname van het aantal electieve operaties, een stijging in het aantal patiënten met lagere NYHA-hartfalen classificaties en een stijging in het aantal patiënten met een recent myocardinfarct.

Voor het laatste hoofdstuk, **Hoofdstuk 10**, ontwikkelden we een getrapte handleiding voor het koppelen van meerdere RCHD-gegevensbronnen op een privacy vriendelijke manier. Met het toenemen van het gebruik van RCHD zal het vaker voorkomen dat men verschillende gegevensbronnen wil koppelen. Het beschermen van de privacy van deelnemers in de gegevensbronnen wordt daarbij steeds belangrijker geacht, iets wat mede te zien is in strenger wordende regelgeving. Om onderzoekers te helpen bij het koppelen van gegevensbronnen op een privacybeschermende manier, hebben wij dit stappenplan voor hen ontwikkeld waarin de belangrijkste stappen voor het koppelen op een privacybeschermende manier van gegevens wordt beschreven.

LIST OF PUBLICATIONS

van der Graaf R, van Dijk W, Laurijssen SJM, Schuit E, Grobbe DE, de Vries MC. The Duty to Support Learning Health Systems: A Broad Rather than a Narrow Interpretation. Am J Bioeth 2021; 21: 14–6.

van Dijk WB, Schuit E. Replacing physical with virtual genetic tests: The importance of conscious methodological decisions. Eur J Prev Cardiol 2020; 27: 1637–8.

van Dijk WB, Grobbee DE. Cardiovascular guidelines based on high-quality evidence: Are we getting there? Eur J Prev Cardiol 2019; 26: 1915–7.

van Dijk WB, Fiolet ATL, Schuit E, et al. Text-mining in electronic healthcare records can be used as efficient tool for screening and data-collection in cardiovascular trials: a multicenter validation study. J Clin Epidemiol 2020; 132: 97–105.

van Dijk WB, Grobbee DE, de Vries MC, Groenwold RHH, van der Graaf R, Schuit E. A systematic breakdown of the levels of evidence supporting the European Society of Cardiology guidelines. Eur J Prev Cardiol 2019; 26: 1944–52.

Vonk ABA, Meesters MI, Van Dijk WB, et al. Ten-year patterns in blood product utilization during cardiothoracic surgery with cardiopulmonary bypass in a tertiary hospital. Transfusion 2014; 54: 2608–16.

DANKWOORD

Prof. dr. D.E. Grobbee, beste Rick, veel respect heb ik voor de ontspannenheid waarmee jij mijn vaak eigenwijze persoon door deze promotie hebt gestuurd. Veel dank voor de begeleiding in de afgelopen jaren en de rol die je hierin hebt gespeeld. Je mij altijd de vrijheid gegeven zelf het pad te kiezen dat ik als bijdragend aan de wetenschap zag. Ook als dit pad (deels) buiten de gebaande paden en projecten ging. Daarnaast heb je me ook kansen geboden door mensen van buiten het Julius Centrum te betrekken en mij de kans te geven me bij projecten aan te sluiten. Hoe jij vrijheid geeft met het op koers houden van een project combineert, benijd ik.

Prof. dr. M.C. de Vries, beste Martine, naast de vorming die jij met Rick aan mijn project gaf, waardeer ik met name de gesprekken die wij hadden over allerhande medisch-ethische onderwerpen. In onze gesprekken stond vaak de zoektocht tussen vooruitgang, goed doen en dit op een juiste manier doen centraal. Je hebt hiermee mijn inhoudelijk denken verder gebracht en bijgedragen aan het succes van veel van mijn projecten. Diep respect heb ik ook voor jouw keuze om je niet alleen in woorden je bezig te houden met goed doen in de wereld, maar dit ook daadwerkelijk omzet in daden in je werk en je avontuur naar het vluchtelingenkamp Moria, in Griekenland.

Dr. E. Schuit en dr. R. van der Graaf, beste Ewoud en Rieke, zonder jullie was dit proefschrift er niet gekomen. Ewoud, jouw rustige aanwezigheid hielp mij als ik weer eens allemaal verschillende richtingen op wilde. Nooit liet jij je van de leg brengen en meestal stuurde je me subtiel bij. Zelden heb ik jou in de spotlight zien staan, maar wat was jij vaak die stille kracht die papers en analyses beter maakte. Rieke, jouw oog voor detail heb ik zeer leren waarderen. Hoewel ik dit zeker niet altijd het leukste vond, heeft het de kwaliteit van dit proefschrift aanzienlijk verhoogd. Ook waardeer ik de gespreken die wij altijd op jouw kamer hadden als ik weer eens langsliep zeer. Je deur stond altijd open.

Leden van de beoordelingscommissie, prof. dr. P.G.M. van der Heijden, prof. dr. L. Hooft, prof. dr. A.W. Hoes, prof. dr. M.M. Levi, prof. dr. M.J. Schalij, graag wil ik u bedanken voor de tijd en moeite die u in het beoordelen van dit proefschrift heeft gestoken.

Prof. dr. R.H.H. Groenwold en dr. A Mosterd, beste Rolf en Arend, beide officieel geen lid van mijn promotieteam, maar officieus ware jullie waarschijnlijk belangrijker voor mij dan jullie zelf wisten. Jullie benaderbaarheid en bereidwilligheid om mij altijd te helpen ben ik zeer dankbaar voor. Niet alleen heeft het dit proefschrift verbeterd, jullie zijn voor mij een voorbeeld voor hoe ik anderen altijd hoop te zullen behandelen.

Msc. (bijna dr.) S.J.M. Laurijssen, beste Sara, als medepromovendus op het ETHMIRE project heb ik veel met je mogen samenwerken en toch ook weinig. Ondanks dat we zo verschillen, gingen we door dezelfde pieken en dalen, en vonden we elkaar in de topics van onze promoties. Ik vond het erg fijn deze met je te kunnen delen. Leuk vond ik ook om te zien hoe jij je passie vond in het onderwijs en dat ook gebruikte als bron van energie voor je onderzoek.

MSc (bijna dr.) A.T.L. Fiolet, beste Aernoud, een maatje kom je niet elke dag tegen, maar heb ik in jou tijdens mijn promotie wel gevonden. Met veel plezier heb ik samen met jou aan onze gezamenlijke papers gewerkt. We vonden elkaar in onze passie om te werken met data uit elektronische patiëntendossiers en vulden elkaar aan in elkanders kwaliteiten. In de toekomst hoop ik nog vaak met je van gedachten te mogen wisselen over onderwerpen die we delen.

Dr. M. Mostert, dr. C.H. Vaartjes, prof. dr. M. Bots, dr. A. Leeuwenberg, Mr. J.A.L Smit, beste Menno, Ilonca, Michiel, Tuur en Julie-Anne, allen hebben jullie meegeschreven aan een of meerdere papers waarop dit proefschrift is gebaseerd. Van recht en epidemiologie tot data science, allen stonden jullie open met mij samen te werken en de breedheid van mijn interesses te ondersteunen. Ik ben blij met jullie gewerkt te hebben.

G. de Vries, beste Giene, met enige regelmaat heb ik mij afgevraagd wie belangrijker was voor het tot stand komen van dit proefschrift: Rick of jij? Hoewel ik niet kan afdoen aan de inhoudelijke hulp van Rick, wil ik jou ook expliciet bedanken voor al jouw hulp bij het tot stand komen van dit proefschrift. Immer was je bereid om mij met vriendelijke daadkracht te helpen bij het navigeren van het Julius Centrum, UMC Utrecht of de ESC.

Dr. Groenhof, beste Katrien, voor niemand anders dan jou geldt zoveel dat dit proefschrift er niet was gekomen zonder jou. Jij wees me op de vacature waarin dit proefschrift geschreven moest worden. Dank daarvoor, ik heb het schrijven van dit proefschrift ervaren als een verrijking. Ook je openheid om altijd mee te denken heb ik altijd zeer gewaardeerd.

Vrienden van De Jonge Dokter, en in het bijzonder Mia, Christiaan, Bas, Meike, Fia en Arno, naast alle tijd die ik in dit proefschrift gestopt heb, heb ik de eerste jaren ook vele uren met jullie doorgebracht om De Jonge Dokter van de grond te krijgen. Niet alleen kijk ik met trots naar waar De Jonge Dokter nu staat, ook ben ik jullie dankbaar voor al jullie reflecties als ik ergens tegenaan liep in mijn onderzoeken.

Vrienden van de Stichting Zorg na Werk in Coronazorg, een geluk bij een ongeluk dat ik jullie heb mogen leren kennen. Het was onfortuinlijk dat de coronaepidemie er tijdens de tweede helft van mijn promotietijd was. Zonder deze pandemie had ik jullie echter nooit leren kennen, en da was een groot gemis geweest. Nog steeds kijk ik met ongeloof terug met ongeloof naar de snelheid waarmee wij deze prachtige stichting hebben kunnen opzetten.

Vrienden van Buurtdokters, en in het bijzonder Bob, Sophie en Sjoerd, een startup kon er ook nog wel bij, en wat ben ik blij dat ik dat met jullie mocht doen. Van de eerste dineravonden in Amsterdam met Bob tot het daadwerkelijk binnenhalen van onze eerste huisartsenpraktijk, ik heb genoten van de actie met jullie in contrast tot het vaak droge academische schrijven. De uren die ik hierover tegen jullie aan heb mogen praten gaven mij de balans met actie die ik nodig had om dit proefschrift te schrijven.

Vrienden van de Nationale DenkTank, mijn deelname voorafgaand aan dit proefschrift in 2016 heeft mijn leven letterlijk van koers veranderd. Vanaf de eerste dag was de brede interesse van jullie allen een warm bad voor mij om in te belanden. Waar ik in de cocon van een ziekenhuis wilde gaan werken, heb ik bij de DenkTank ingezien liever naar wisselende onderwerpen vanaf een ander niveau naar de maatschappij te kijken. Iets wat ten grondslag ligt aan mijn keuze voor de consultancy nu.

“Vrienden van vroeger en nu”, Sjaak en Hayk, ook jullie verdienen een plaats hier in dit dankwoord. Hoewel jullie beiden niet direct met dit boekje van doen hadden, waardeer ik het luisterend oor wat jullie mij altijd boden als ik vaak met enthousiasme over een voor jullie vermoedelijk niet te volgen onderwerp aan het ratelen was.

Chris, Lot en Joost, woorden zullen er nooit genoeg zijn om jullie te bedanken voor de steun die jullie mij altijd bieden. Jullie hebben het tot stand komen van dit proefschrift op afstand mogen aanschouwen en mijn geleuter over te academische onderwerpen moeten aanhoren. Dank voor jullie altijd aanwezige steun. Ik houd van jullie

Lieve Ursje, ik ben erg blij dat ik mijn promotietraject de afgelopen jaren met jou heb mogen delen. Meer dan wie dan ook heb jij mijn verhalen aan moeten horen. Van ’s ochtends vroeg tot ’s avonds laat, doordeweeks en in het weekend, je weet inmiddels: mijn hoofd staat nooit stil. En hoewel je mijn gewauwel niet op elk moment kon waarderen, luisterde je wel altijd en kwam je met inzichten waar ik zelf niet op gekomen zou zijn. Ik ben blij samen met jou (en sinds kort Jules) te mogen zijn. Je weet: ik houd van je.

ABOUT THE AUTHOR

Wouter van Dijk was born on December 27, 1988 in Amsterdam and raised in Almere. In 2009, Wouter passed his secondary school exam at the Helen Parkhurst in Almere.

Wouter studied medicine at the University of Utrecht (2009-2016, BSc, MSc, MD) and law at the University of Amsterdam (2012-2018, LLB, LLM). During his study, he participated in multiple committees addressing the professional development of future doctors in medical studies. Wouter also followed two of his rotations aboard, one in León, Nicaragua (gynecology) and one in Machame, Tanzania (public health). In his final year, Wouter participated in the Vital Signs excellence program, focusing to develop himself in the field of anesthesiology.

After his studies, Wouter participated as a participant in the Dutch National ThinkTank (2016). In the National ThinkTank Wouter and 20 other recent graduates were challenged to research and investigate solutions for the Dutch vocational education system. After his participation in the Dutch National ThinkTank, Wouter briefly joined McKinsey and Company as a healthcare consultant (2017) before starting to work on his PhD.

During the work on his PhD (2018-2021), Wouter participated in multiple conferences, both with poster and oral presentations. Moreover, Wouter was one of the five nominees for the European Society of Cardiology Young Investigator Awards in 2020. While working on his PhD, Wouter founded the Dutch Young Doctors Foundation and Society (2018), the Foundation Care After Work in Corona care (2020) and a healthcare general practice chain start-up, Buurtdokters (2020).