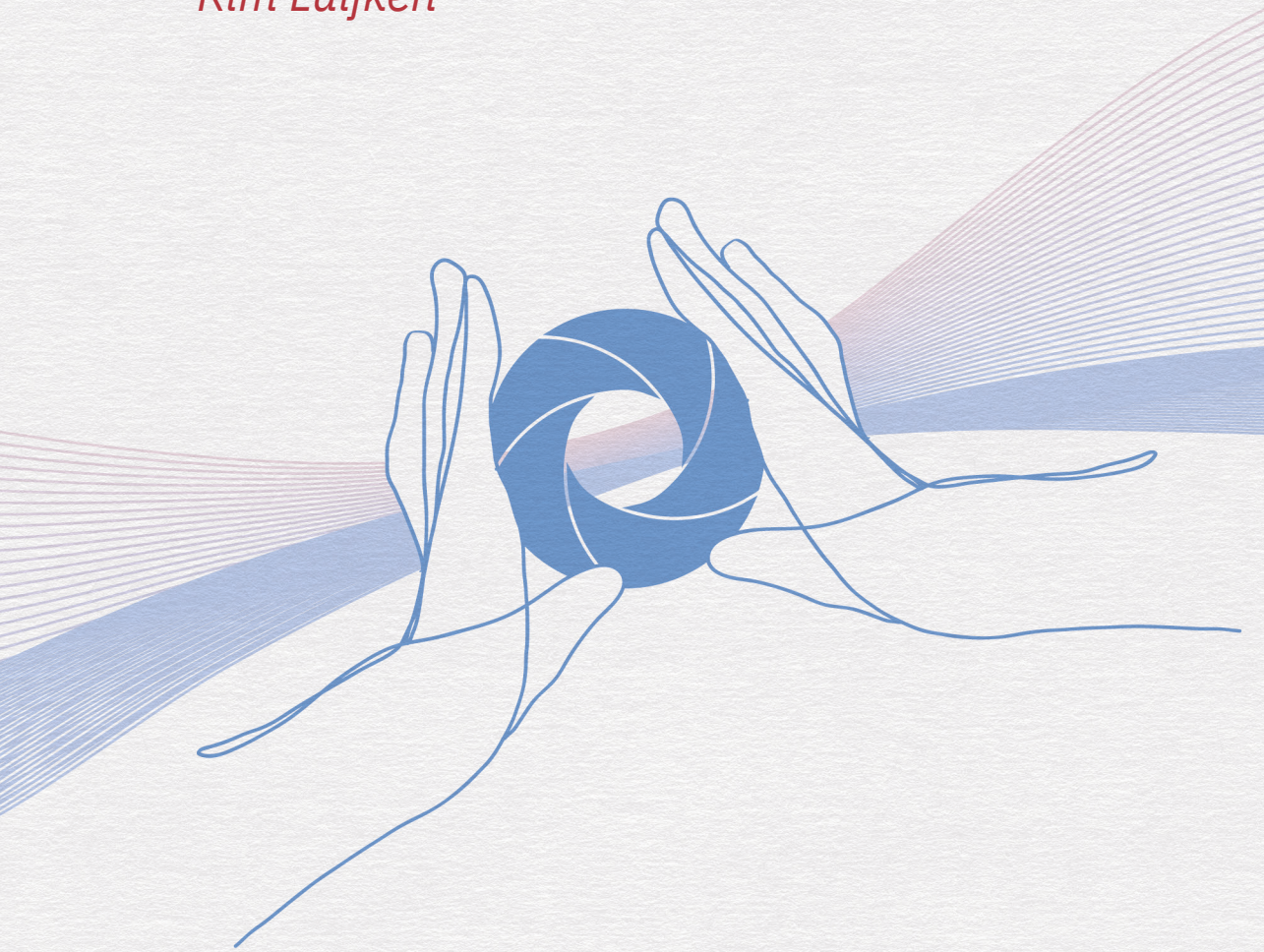
The impact of epidemiologic methods on findings in studies of causal effects and prediction modelling

Luijken, K.

The impact of epidemiologic methods on findings in studies of causal effects and prediction modelling

Kim Luijken



The impact of epidemiologic methods on findings in studies of causal effects and prediction modelling

Kim Luijken

The impact of epidemiologic methods on findings in studies of causal effects and prediction modelling

PhD thesis. Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, the Netherlands

ISBN: 978-94-6416-842-6
Author: Kim Luijken
Printing: Ridderprint, www.ridderprint.nl
Layout and Cover Design: Harma Makken, persoonlijkproefschrift.nl

The impact of epidemiologic methods on findings in studies of causal effects and prediction modelling

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Leiden
op gezag van de rector magnificus Prof. dr. ir. H. Bijl,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 19 mei 2022
klokke 13:45 uur

door
Kim Luijken

Geboren op 26 april 1994
te Rotterdam

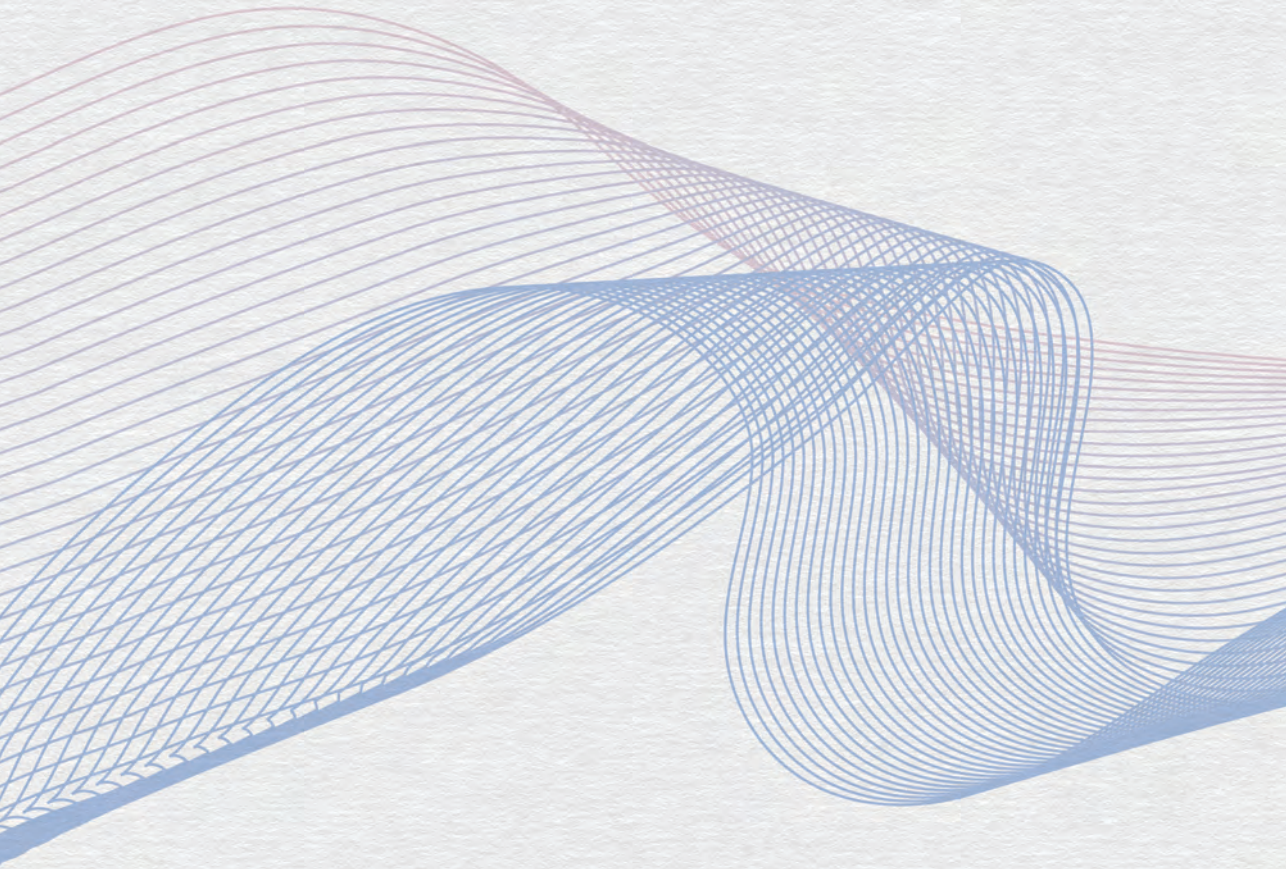
Promotor:	Prof. dr. R.H.H. Groenwold	
Promotiecommissie:	Prof. dr. O.M. Dekkers	
	Prof. dr. M. Fiocco	
	Prof. dr. E.W. Steyerberg	
	Prof. dr. I.G. Klugkist	Universiteit Utrecht
	Prof. dr. L. Hooft	Universitair Medisch Centrum Utrecht

"Methodology almost never perfectly corresponds to the complex phenomena that give rise to our data. Methodology within a field ought to advance in expanding the range of questions that can be addressed, in relaxing the assumptions required, and in allowing investigators to assess the sensitivity of conclusions to violations in the assumptions."

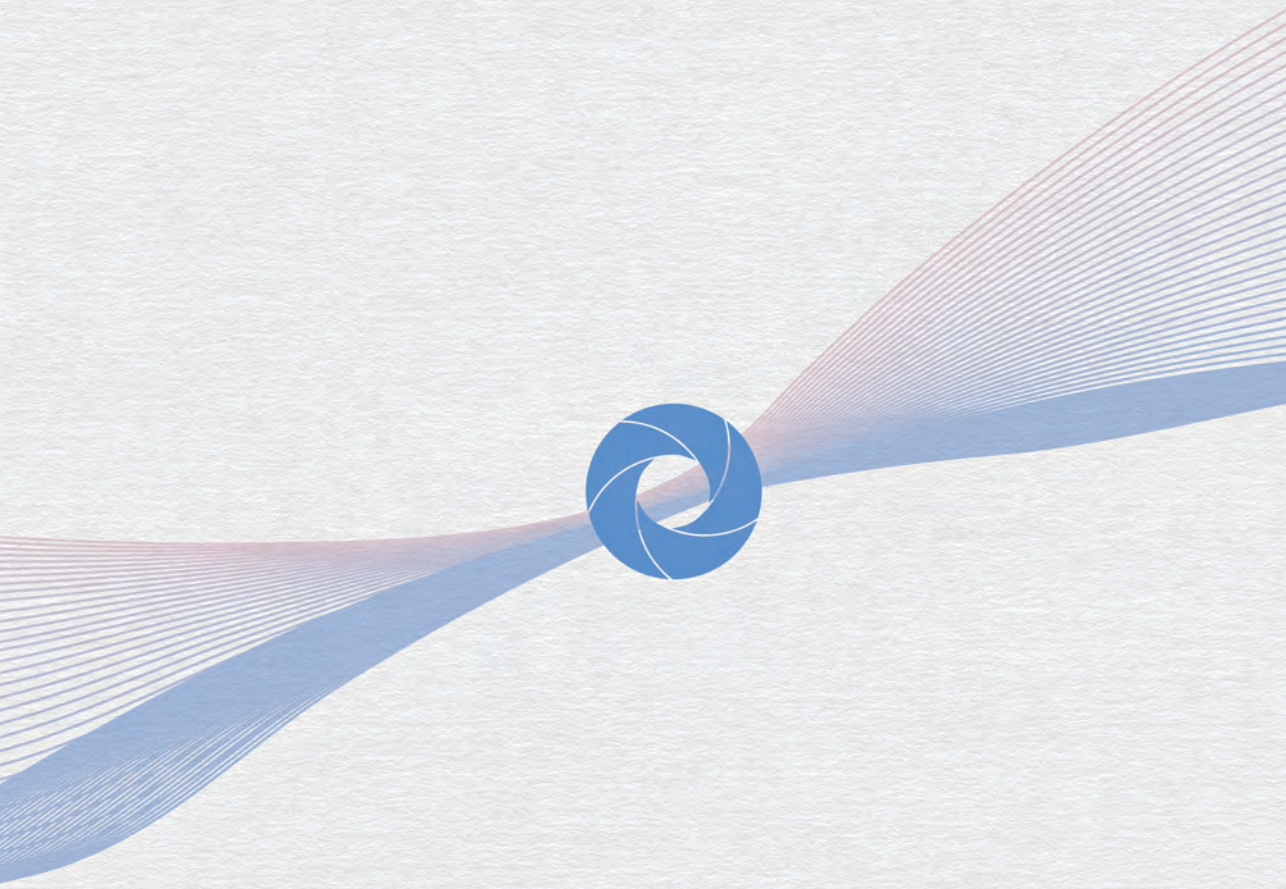
VanderWeele TJ, Vansteelandt S. Respond to "Decomposing With a Lot of Supposing" and "Mediation", *American Journal of Epidemiology*. 2010;172(12):1355–1356.

Content

Part II: impact of applied methods on the meaning of numeric estimates in prediction modelling studies



1



General introduction

Developments in epidemiological practice

Quantitative scientific evidence is a key pillar of current evidence-based medicine to inform medical decision making[1]. To aid understanding of clinical questions through numerical results, quantitative approaches are continuously being developed in the field of clinical epidemiology and medical statistics, hereafter referred to as the field of 'epidemiology'. During most of the 20th century, study designs mainly revolved around operationalizations of the exposure and outcome variables, while statistical approaches focused on models that required no or minimal modelling assumptions[2]. Epidemiological approaches were used to address relatively broad, yet fundamental clinical questions, such as: 'does smoking cause lung cancer?'. As an example, this question was addressed in one of the first observational clinical studies, performed by the Sirs Doll and Hill[3]. It can be computed from their results that the odds of developing lung cancer when smoking was around 14 times the odds of developing lung cancer when not smoking: a substantial effect size.

Since then, the apparatus of epidemiological study designs and data analytical approaches has expanded considerably. The latter half of the twentieth century saw a strong rise of epidemiological evidence as a basis of medical decision making, where the hierarchy of evidence from clinical experience to epidemiological evidence essentially reversed[4,5]. More data sources became available for research purposes, including routinely collected clinical data[6]. Since the increase in data volume was not always paralleled by data quality[7], study designs and methods were increasingly refined to address variation in data collection procedures. Statistical techniques were developed to accommodate the higher complexity of the data, from multivariable models to machine learning, aided by an extraordinary increase in computing power. These refined epidemiological approaches are applied in a wide variety of public health and medical research fields to address specific questions, such as: 'is cotinine level (a biomarker for passive tobacco smoke exposure) associated with an increased 20-year risk of coronary heart disease in men who self-report to be non-smokers?'. A study by Whincup and colleagues found that the hazard of developing coronary heart disease over 20 years among non-smoking men with a cotinine concentration of 2.8 – 14.0 ng/mL was 1.57 times the hazard of developing coronary heart disease among men with a cotinine concentration of ≤ 0.7 ng/mL, conditional on established risk factors for coronary heart disease (with a 95% confidence interval from 1.08 to 2.28)[8]. Compared to the marginal odds ratio of 14 found by Doll and Hill, this association appears to be

weak, meaning that methodological and statistical artefacts increasingly run the risk of bearing consequences for interpretation of findings.

The seemingly straightforward methodological principles of early epidemiological research and the current abundancy in methods are contrasted here to outline that the increasingly central position of formalized quantitative methods requires careful consideration of the analytical choices to be made during study conduct.

Analytical tools at the disposal of clinical questions

It is a challenge to keep abreast of all developments in epidemiological methods. To facilitate application of statistical approaches, a large number of tutorial series and coding vignettes break down statistical material into manageable endeavors (for instance, the 'Statistical Notes' series in the *British Medical Journal*[9] or vignettes of the R package epiR[10]). Some modelling procedures are even fully automated, such as for instance a high-dimensional propensity score algorithm for analysis of causal effects of pharmacological treatments in routinely collected health care databases[11,12], or software that automates prognostic modelling in observational databases using the Observational Medical Outcomes Partnership (OMOP) Common Data Model format[13], to the degree that manuscript output can be partly rendered by the statistical package[14]. These tutorials and packages ease the use of statistical tools, but the (default) method that is applied may not be optimal for a particular study[15]. Moreover, the choice of the analytical method directly influences the clinical interpretation of the results of the analysis and thus research findings, yet it is undesirable that technical decisions define the subject of the investigation.

The challenge of identifying a study design and statistical method for data analysis that appropriately helps answering the research question of clinical interest has long been recognized in epidemiological research. At the basis of epidemiological research is a careful definition of the research question and a plan how to conduct the study accordingly (Figure 1). Many epidemiology textbooks start from the premise that study conduct should be informed by a carefully defined research question, e.g., [16,17]. Statistical literature increasingly emphasizes that the study aim – identifying it as descriptive, predictive, or causal – is the basis for choosing an appropriate statistical model and interpreting the findings correctly[18-20]. An addendum to the ICH E9 guideline Statistical Principles for Clinical trials further explicated the link between clinical aims and analytical theory by clarifying what the specification of a study estimand entails[21].

Defining the research question and corresponding estimand provides a link between the study aims and its analysis and endows the numerical result with a clinically meaningful interpretation.

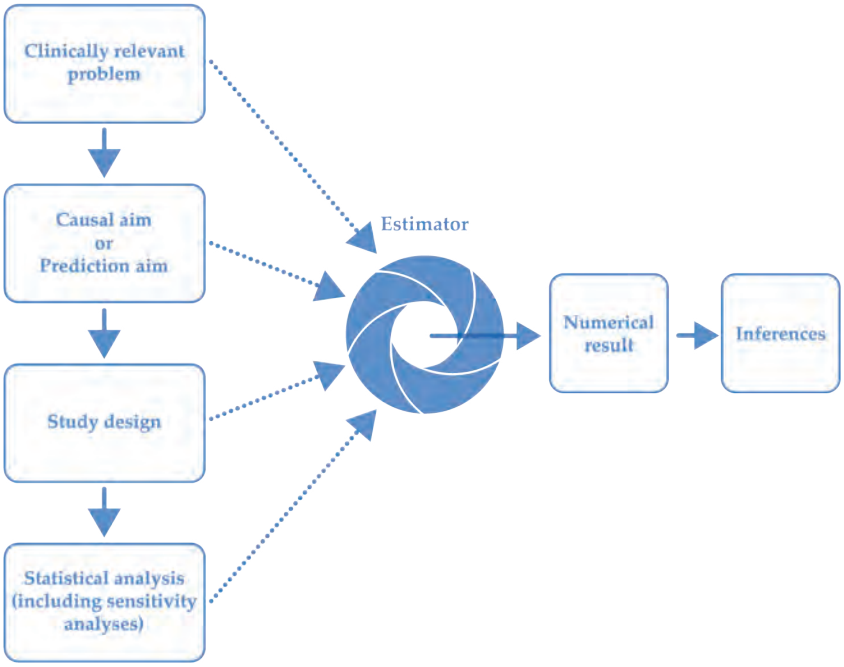


Figure 1. Schematic depiction of stages of research conduct of a quantitative study. The estimator is depicted as a diaphragm, resembling the aperture around the lens through which a research problem is studied in a quantitative analysis. The way the diaphragm is tuned by the clinical context, study aim, study design, and statistical analysis plan determines what can be observed through the lens and thus directly affects inferences that can be made from the estimate. A scheme like this should be accompanied by the remark that "Trying to cash out a full-blown picture of inquiry that purports to represent all contexts of inquiry is a fool's errand. […] If one is not to land in a Rube Goldberg mess of arrows and boxes, only to discover it's not pertinent to every inquiry, it's best to settle for pigeonholes roomy enough to organize the interconnected pieces of a given inquiry" [22, p. 86-7].

Challenges in generating clinically meaningful estimates

Defining an estimand and choosing an adequate study design and statistical analysis that align with the clinical aim of a study is clear in principle yet can be complicated in practice. The clinical interpretation of estimates has been clearly described for some estimators[23-28], but the link between technical procedures and substantive interpretation sometimes remains implicit in methodological guidance. As an illustration, most risk of bias tools that are widely implemented to assess the validity of a study do not explicitly address how to assess the potential for bias with respect to the aim of a particular study[29-32].

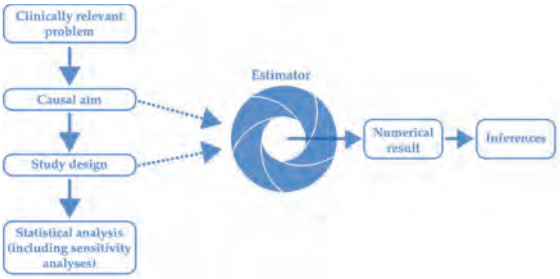
Having a deeper understanding of the impact that data analytical decisions can have on the interpretation of numerical results of a study would help to use methodology as a means to a clinical end. Comprehension of the influence that design choices might have on interpretation of estimates would be insightful for etiologic research and prediction research in particular, since most discussions on estimands and alignment with study conduct to date focused on therapeutic research.

Thesis aim and outline

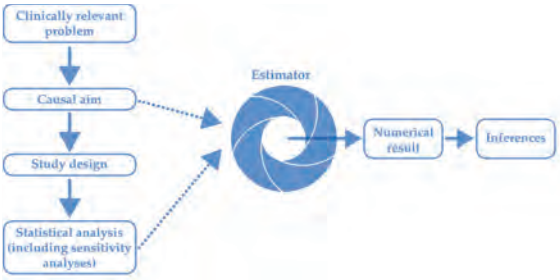
As described above, epidemiological methods have been increasingly refined, but it may not always be clear what the impact is of (habitual) choices regarding the design and statistical analysis of a study on the meaning of its numerical results. The aim of this thesis is to investigate this impact, where we focus separately on research into causal effects (Chapters 2-5) and prediction research (Chapters 6-8).

Part I: impact of applied methods on the meaning of numeric estimates in studies of causal effects

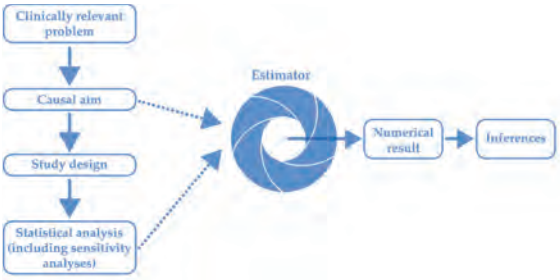
The first set of case studies (Chapter 2 – 5) focuses on causal research.



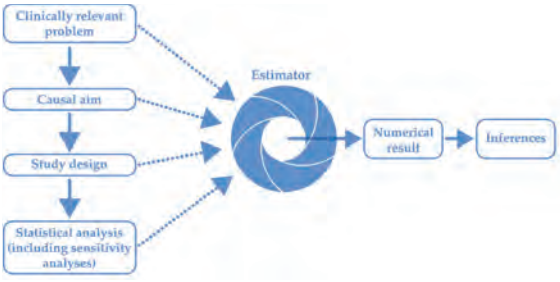
Chapter 2 evaluates how the time origin of a study design affects the interpretation of numerical results through a systematic review of reporting practices of the estimand of interest and the definition of study time origin in observational comparative effectiveness and safety cohort studies into effects of pharmacological treatments.



Chapter 3 discusses the implications of being ambiguous about the study aim in exploratory etiologic studies. We define a continuum of scrutiny with which exploratory statistical analyses are conducted and provide practical pointers for good practice in exploratory etiological research.



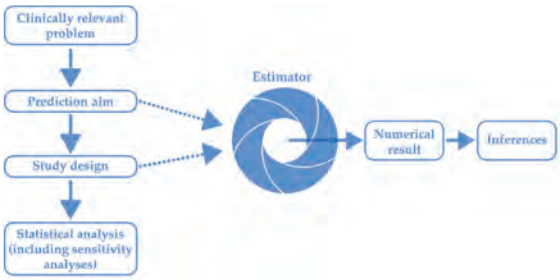
Chapter 4 evaluates the impact of the choice of a statistical approach on interpretation of results, by investigating under which conditions causal effect estimation in observational studies improves by using backward elimination on a prespecified set of potential confounders.



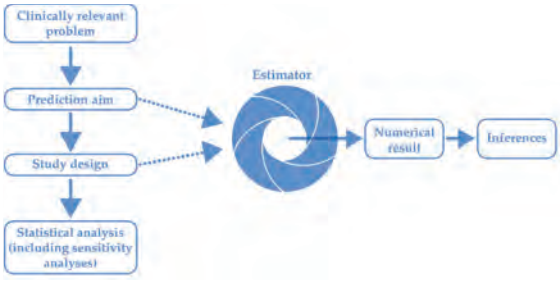
Chapter 5 describes how the link between the research question and study conduct can be assessed in studies of operative interventions, for instance as part of a systematic review. We propose an easy-to-use concise set of items derived from existing tools that is intended to perform an initial assessment of the applicability and methodological quality of research into the effects of surgical interventions.

Part II: impact of applied methods on the meaning of numeric estimates in prediction modelling studies

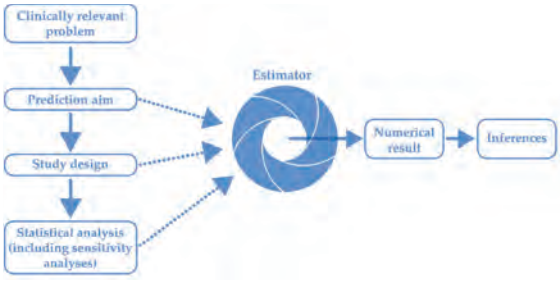
Chapters 6 – 8 form an in-depth case study in prediction research evaluating the impact of predictor measurement procedures on the performance of prediction models.



Chapter 6 uses an established taxonomy of measurement error models to define and clarify *predictor measurement heterogeneity*: differences in predictor measurement strategies across settings of derivation and validation. It is investigated whether predictor measurement heterogeneity affects the predictive performance of binary logistic prediction models, using analytical expressions and simulations.



Chapter 7 provides an illustration of the impact of predictor measurement heterogeneity on predictive performance in clinical examples using previously developed prediction models for diagnosis of ovarian cancer, mutation carriers for Lynch syndrome, and intrauterine pregnancy.



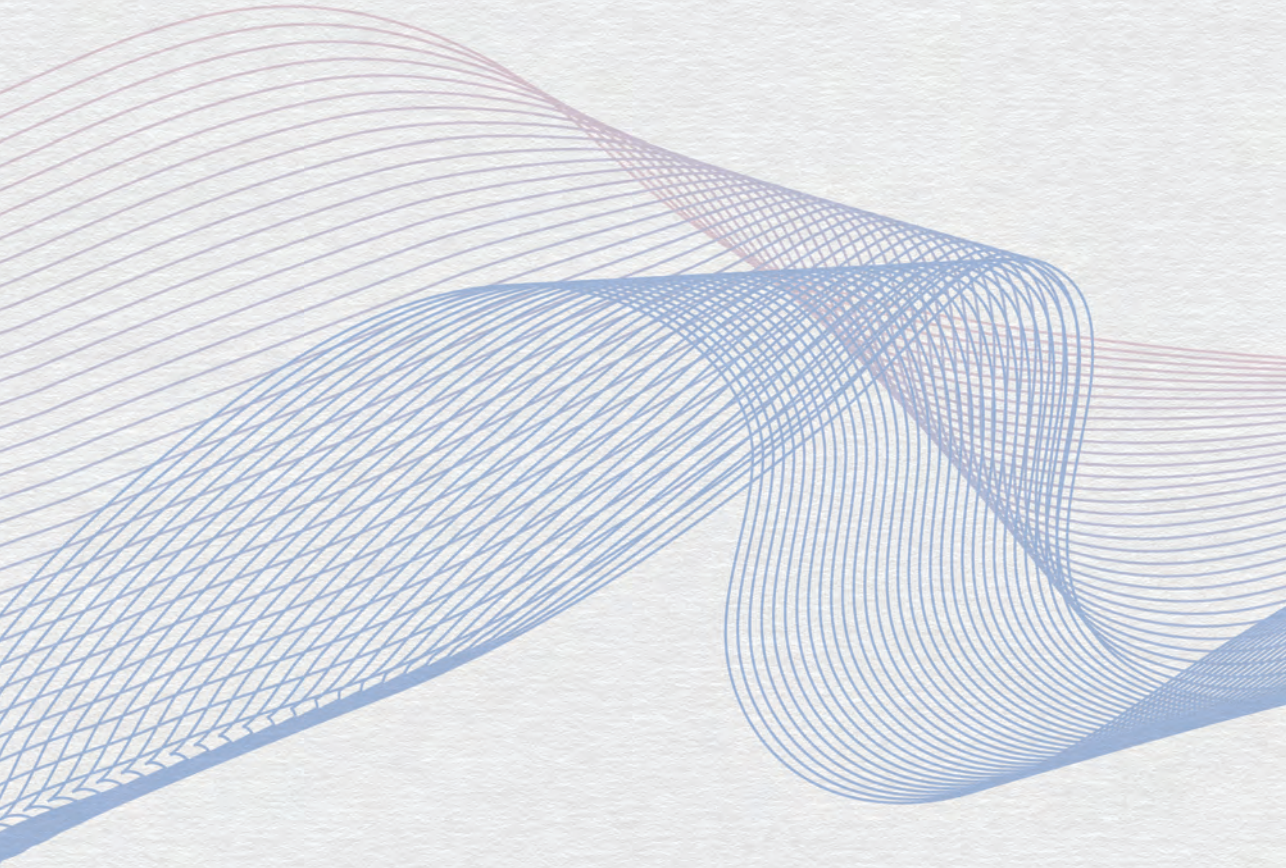
Chapter 8 presents a quantitative prediction analysis that can be used in validation studies to anticipate the impact of predictor measurement heterogeneity on predictive performance of time-to-event prediction models at implementation in practice if one of the predictor measurements is expected to deviate from the prediction target.

Chapter 9 summarizes the impact of applied methodology on interpretation of results described in these examples and ends with a general discussion on strengthening research question formulation in clinical studies.

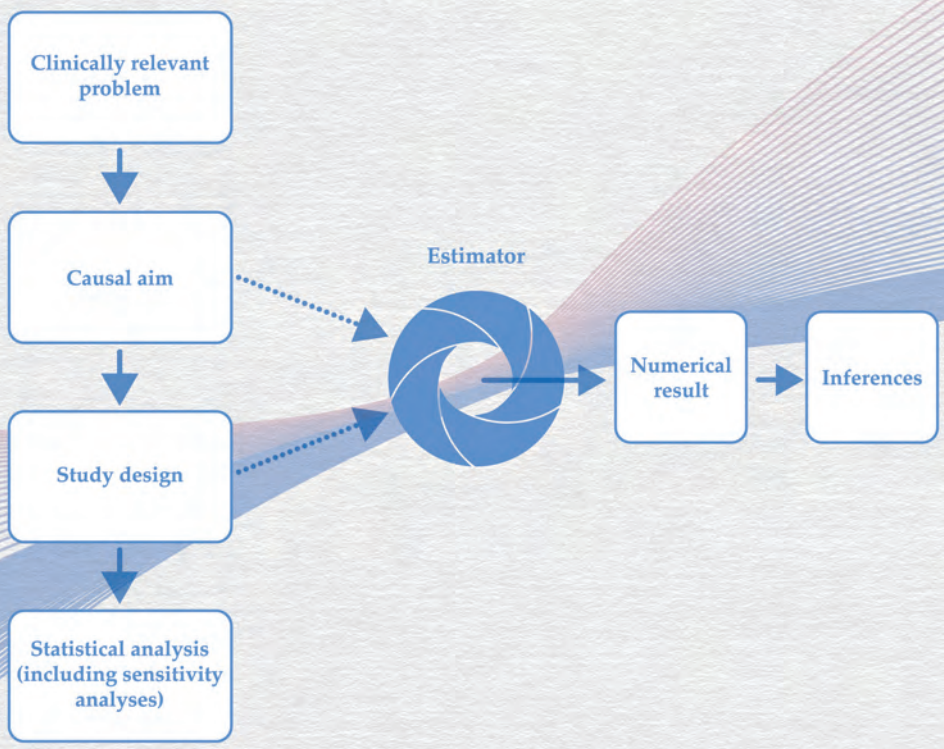
References

1. Dickersin K, Straus SE, Bero LA. Evidence based medicine: increasing, not dictating, choice. *British Medical Journal*. 2007;334:s10.
2. Hill AB. *Principles of Medical Statistics*. London: Lancet; 1937.
3. Doll R, Hill AB. Smoking and carcinoma of the lung. *British Medical Journal*. 1950;2(4682):739.
4. Guyatt G, Cairns J, Churchill D, et al. Evidence-based medicine: a new approach to teaching the practice of medicine. *Journal of the American Medical Association*. 1992;268(17):2420-2425.
5. Porter TM. *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press; 2020.
6. Mooney SJ, Westreich DJ, El-Sayed AM. Epidemiology in the era of big data. *Epidemiology (Cambridge, Mass)*. 2015;26(3):390.
7. Hand DJ. Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2018;181(3):555-605.
8. Whincup PH, Gilg JA, Emberson JR, et al. Passive smoking and risk of coronary heart disease and stroke: prospective study with cotinine measurement. *British Medical Journal*. 2004;329(7459):200-205.
9. Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar VS, Grimmer KA. A systematic review of the content of critical appraisal tools. *BMC Medical Research Methodology*. 2004;4(1):1-11.
10. Stevenson M. Sample Size Calculations Using epiR. https://cran.r-project.org/web/packages/epiR/vignettes/epiR_sample_size.html. Published 2021. Accessed 28-05-2021.
11. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass)*. 2009;20(4):512.
12. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clinical Epidemiology*. 2018;10:771.
13. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Annals of Internal Medicine*. 2010;153(9):600-606.
14. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association*. 2018;25(8):969-975.
15. Feinstein AR. "Clinical judgment" revisited: the distraction of quantitative models. *Annals of Internal Medicine*. 1994;120(9):799-805.
16. Westreich D. *Epidemiology by design: a causal approach to the health sciences*. Oxford University Press; 2019.
17. Lash TL, VanderWeele TJ, Haneause S, Rothman K. *Modern epidemiology*. Lippincott Williams & Wilkins; 2020.
18. Shmueli G. To explain or to predict? *Statistical Science*. 2010;25(3):289-310.
19. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *Chance*. 2019;32(1):42-49.
20. Van der Laan MJ, Rose S. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media; 2011.
21. ICH E9 working group. ICH E9 (R1): addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Published 2020. Accessed 28-05-2021.
22. Mayo DG. *Statistical inference as severe testing*. Cambridge University Press; 2018.
23. Ratitch B, Bell J, Mallinckrodt C, et al. Choosing estimands in clinical trials: putting the ICH E9 (R1) into practice. *Therapeutic Innovation & Regulatory Science*. 2020;54(2):324-341.

24. Mallinckrodt C, Bell J, Liu G, et al. Aligning estimators with estimands in clinical trials: putting the ICH E9 (R1) guidelines into practice. *Therapeutic Innovation & Regulatory Science.* 2020;54(2):353-364.
25. Ratitch B, Goel N, Mallinckrodt C, et al. Defining efficacy estimands in clinical trials: examples illustrating ICH E9 (R1) guidelines. *Therapeutic Innovation & Regulatory Science.* 2020;54(2):370-384.
26. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology (Cambridge, Mass).* 2014;25(3):418.
27. Greifer N, Stuart EA. Choosing the Estimand When Matching or Weighting in Observational Studies. *arXiv preprint arXiv:210610577.* 2021.
28. Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I, initiative ttgCIotS. Formulating causal questions and principled statistical answers. *Statistics in Medicine.* 2020;39(30):4922-4948.
29. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine.* 2011;155(8):529-536.
30. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. *Annals of Internal Medicine.* 2013;158(4):280-286.
31. Higgins JP, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal.* 2011;343.
32. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Published 2000. Accessed 28-05-2021.



2


Clinically relevant problem
Causal aim
Study design
Statistical analysis (including sensitivity analyses)
Estimator
Numerical result
Inferences

New-user and prevalent-user designs and the definition of study time origin in pharmacoepidemiology: a review of reporting practices

Guidance reports for observational comparative effectiveness and drug safety research recommend implementing a new-user design whenever possible, since it reduces the risk of selection bias in exposure effect estimation compared to a prevalent-user design. The uptake of this guidance has not been studied extensively. We reviewed 89 observational effectiveness and safety cohort studies published in six pharmacoepidemiologic journals in 2018 and 2019. We developed an extraction tool to assess how frequently new-user and prevalent-user designs were reported to be implemented. For studies that implemented a new-user design in both exposure arms, we extracted information about the extent to which the moment of meeting eligibility criteria, treatment initiation, and start of follow-up were reported to be aligned. Of the 89 studies included, 40% reported implementing a new-user design for both the study exposure arm and the comparator arm, while 13% reported implementing a prevalent-user design in both arms. The moment of meeting eligibility criteria, treatment initiation, and start of follow-up were reported to be aligned in both exposure arms in 53% of studies that reported implementing a new-user design. We provided examples of studies that minimized the risk of introducing bias due to unclear definition of time origin in unexposed participants, immortal time, or a time lag. To sum up, almost half of the included studies reported to implement a new-user design. Implications of misalignment of study design origin were difficult to assess because it would require explicit reporting of the target causal effect in original studies. We recommend that the choice for a particular study time origin is explicitly motivated to enable assessment of validity of the study.

This chapter was based on: Luijken K, Spekreijse JJ, van Smeden M, Gardarsdottir H, Groenwold RHH. New-user and prevalent-user designs and the definition of study time origin in pharmacoepidemiology: A review of reporting practices. *Pharmacoepidemiology and Drug Safety*. 2020;30(7):960-74.

1 | Background

Guidance reports for comparative effectiveness and safety research of pharmacological treatments recommend the new-user design[1-4], in which follow-up time generally starts with the first prescription or dispensing of the drug(s) of interest[5]. In contrast, in the prevalent-user design both current (prevalent) and new users of a drug are included. The new-user design enforces appropriate temporal ordering of measurements of confounders, treatment, and outcome, protecting the researcher against accidental adjustment for variables affected by treatment and against finding associations that are based on reversed causation[1-8]. However, the start of a treatment can be difficult to capture (especially in case of intermittently used treatments) and exclusion of prevalent users may reduce follow-up time or sample size[5,7-10]. It is unclear how often and for which reasons researchers deviated from the guidance to implement a new-user design.

To assess the uptake of new-user design guidance, it is important to go beyond the distinction of including new or prevalent users. Many time-related biases can be prevented by choosing a study time origin (or study baseline) such that it establishes alignment of the moment of meeting eligibility criteria, treatment initiation, and start of follow-up[6,11-13]. Previous studies investigated how often pharmacoepidemiologic studies deviated from the recommendation to implement a new-user design[14-16], however, the implementation of new-user designs in terms of alignment of eligibility, treatment initiation, and start of follow-up has not been studied yet.

In the current study, we reviewed the literature about contemporary observational effectiveness and safety cohort studies. We assessed how frequently new-user and prevalent-user designs were reported to be implemented in studies published in high-ranked pharmacoepidemiologic journals. For studies implementing a new-user design, we evaluated to what extent eligibility, treatment initiation, and start of follow-up were reported to be aligned.

2 | Defining a study time origin

Here, we briefly review common biases that can be introduced by inappropriate designation of follow-up time in observational studies (for a more elaborate discussion, see e.g.,[3,5-7,11,12,17-19]). Ideally, a study is designed such that operational decisions match the specified causal contrast of interest, i.e., the target causal effect or so-called estimand. Whether the target causal effect can be estimated from the available data, depends on

whether it is convincing that untestable identifying assumptions hold, i.e., conditional exchangeability, positivity and consistency. The choice of the time origin of a study design is directly linked to the target causal effect, because the estimated outcome risk refers to the (cumulative) probability of an event of interest occurring over time since a given origin in a specific population[18].

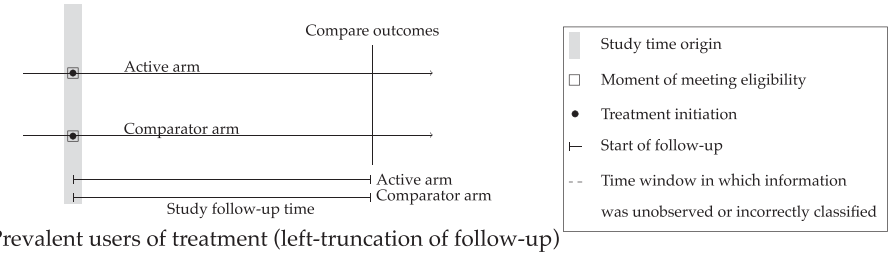
Figure 1 shows a simplified version of a study of a binary treatment in which eligibility, treatment initiation, and start of follow-up correspond[4]. An archetypical study design is the active-comparator incident-user design, that potentially allows identification of the target causal effect from empirical data. When an appropriate active comparator is chosen, in the sense that the comparison group reflects a clinically meaningful alternative treatment option in real-world practice, the active-comparator incident-user design increases the likelihood of achieving conditional exchangeability by minimizing the risk of confounding by indication and selection bias[20].

When prevalent users of treatment are included in a study, the follow-up of those individuals is left truncated and omitted from the analysis. For permanent outcomes, of which 'death' is arguably the clearest example, individuals included in the analysis did not develop the outcome during the exposed period before start of follow-up. Consequently, including prevalent users can lead to under-ascertainment of events early in the course of treatment and increases the risk of controlling for confounding factors that were affected by the treatment[5,21]. Since much information is unobserved, it will be difficult to specify the set of covariates that are sufficient to achieve conditional exchangeability of treatment groups. In many cases, it is unlikely that a prevalent-user study design is suitable to identify the specified target causal effect (Figure 1b).

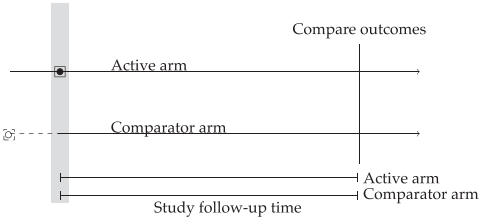
Another issue is that individuals could be assigned to a treatment group based on the treatment strategy observed after start of follow-up, rather than the treatment strategy at the time of start of follow-up[11] (Figure 1c), for example when individuals are classified as 'users' only after they filled a number (e.g., three) of prescriptions of that treatment. In that case, individuals cannot experience the outcome during the first three prescriptions. This period is often referred to as *immortal time*. When some individuals are immortal for part of the time they were followed-up, it seems unlikely that the target causal effect can be identified.

A final example of a study design that can impede identification of the target causal effect is when a time-lag bias is introduced. This happens when follow-up is started at the moment of treatment initiation, but the compared treatments are prescribed at different stages of the disease[13]. For instance, if the effect of a second-line drug

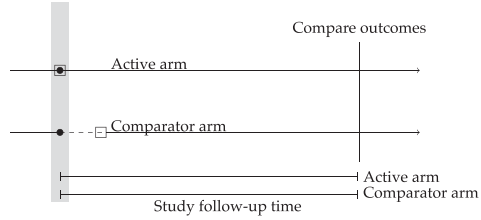
(a) Ideal causal contrast



(b) Prevalent users of treatment (left-truncation of follow-up)



(c) Follow-up time incorrectly allocated (immortal time)



(d) Time-lag in start of follow-up (inexchangeability)

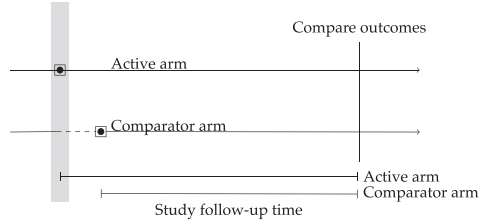


Figure 1. Schematic depiction of possible operationalizations of a study time origin. Subfigure (a) depicts a simplified ideal causal contrast for a binary treatment. Subfigures (b) - (d) depict possible biases that can be introduced by inappropriate designation of follow-up time. In empirical studies, the specified design flaws can occur in either or both exposure arms and a combination of flaws can occur. For simplification, we presented a single design flaw in the comparator exposure arm in each subfigure (b) - (d). Study design flaws may lead to violation of identifying assumptions, as is explained in Section 2 of the main text. The dotted lines in the figure thus indicate the consideration whether exchangeability of treatment groups is jeopardized by the misalignment and whether this can be corrected by measured covariates, or non-positivity is introduced.

is compared to a first-line drug and the start of follow-up for both the active and comparator arm is defined by treatment initiation, the disease stage differs between the treatment groups. As capturing this difference in disease progress in measured covariates is hardly feasible[22], time-lag bias likely jeopardises the (conditional) exchangeability of treatment groups (Figure 1d). When protocol adherence to switch to second line is high and precise, the incomparability of treatment groups may be so extreme that non-positivity is introduced.

3 | Methods

We systematically assessed the reporting practices in observational studies of treatment effects regarding the definition of the study time origin and inclusion of new versus prevalent users of treatment. A protocol of this study is available on Open Science Framework[23]. Based on recommendations by the editor and reviewers, we deviated from this protocol. Specifically, while we scored the items of the extraction tool for all included articles, we discuss the results on alignment in study design origin for new-user designs only, as will be explained below. This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[24], where applicable.

3.1 | Journal selection and included type of studies

We aimed to review the reporting of approximately 100 articles published before the 1st of July 2019 in journals publishing pharmacoepidemiologic studies of drug-outcome associations. Six pharmacoepidemiologic journals were included: Annals of Pharmacotherapy, British Journal of Clinical Pharmacology, Drug Safety, European Journal of Clinical Pharmacology, Pharmacotherapy, and Pharmacoepidemiology and Drug Safety. These state-of-the art pharmacoepidemiologic journals were selected because reporting on study design implementation was expected to be relatively complete. We performed a PubMed search on February 3rd 2020 (see protocol[23] for search string) which returned 2,457 records. Study inclusion criteria were: study described original pharmacoepidemiologic research into the relation between drug exposure and a clinical outcome; data were collected for research purposes or obtained from routinely collected health data; data were gathered according to a cohort study design, since the definition of new versus prevalent users is not as straightforward in other designs, such as a cross-sectional, case-crossover or case-control design. Exclusion

criteria were: pharmacokinetic-pharmacodynamic studies; cost-effectiveness studies; data on treatment exposure were collected through self-report. We also excluded studies of vaccination, antibiotic treatment of a single treatment episode (up to 10 days), chemotherapy, or intravenous drugs, because for these kinds of interventions new-user designs are more natural. KL screened the title and abstract of all studies that result from the searches and included relevant articles based on the eligibility criteria. We applied a quota sampling strategy[25] and continued screening articles until we reached the most recent 100 articles published before July 1st, 2019.

3.2 | Extraction of study characteristics and evaluation of reporting quality

Articles were scored on a set of items derived from guideline recommendations about elements that should be reported in protocols[20,26] or articles[4,27] of effectiveness and safety research using large observational databases, as well as methodological articles that discuss the study time origin in observational studies of causal effects[6,11]. The main focus was on the distinction between new-user and prevalent-user designs and the alignment of moment of meeting eligibility criteria, moment of treatment initiation, and start of follow-up in new-user designs. The established scoring tool was pilot tested on six randomly chosen included studies by KL and JS and further adjusted (all items can be found in Table 2 and 3).

An incident user can more generally be defined as a new user of any treatment decision, i.e., initiating a treatment, but also switching to a different treatment or a change of dose. This understanding of the incident-user design was introduced by Brookhart[28] and expanded in work by Suissa[29] and will be used throughout the current study. For the item that scored reporting of whether the comparator exposure arm implemented a new-user or prevalent-user design, we decided to score nonusers of treatment as prevalent users. Whereas non-use is not associated with the biases typically associated with prevalent users (e.g., adjusting for intermediates, depletion of susceptibles), definition of study time origin in studies with a non-user comparator arm is complicated because the choice of the time origin since which the (cumulative) probability of an event of interest can occur in the specified population may not be as straightforward for non-users of treatment. Consequently, it is more challenging to assess whether the study exposure arm and comparator arm can be assumed to be comparable conditional on measured confounders (i.e., whether there is conditional exchangeability).

Information was gathered on general characteristics of the included studies; funding source, type of data source, patient domain, sample size, and length of enrollment window. Funding source was defined as 'private' when the article stated the study

was funded by a pharmaceutical company or when any of the authors was affiliated with a pharmaceutical company and defined as 'public' otherwise. Data sources were classified into hospital data, dispensings, prescriptions, or claims. Patient domain was grouped into medical specialties based on the target population that was mentioned in the article objective. When the target population did not match a single medical specialty, information on the type of treatment and study outcome was used to identify the medical specialty.

Articles were reviewed independently by KL and JS, results were discussed between the two reviewers and in case of disagreement a third reviewer (RG) was consulted. When multiple effectiveness or safety analyses were described in a single article, only the first-reported analysis was scored. When subgroup analyses were performed in the included studies, only the main analysis was scored. When methods were discussed in an online protocol or described in a different article, we reviewed the referred material.

3.3 | Data synthesis

Rater agreement was computed using the unweighted Cohen's kappa for nominal variables and two coders[30]. Cohen's kappa ranges from -1 (perfect disagreement) to 1 (perfect agreement). Reporting of items was presented as percentages of total number of included studies and 95% confidence intervals (CIs) were computed using the normal approximation.

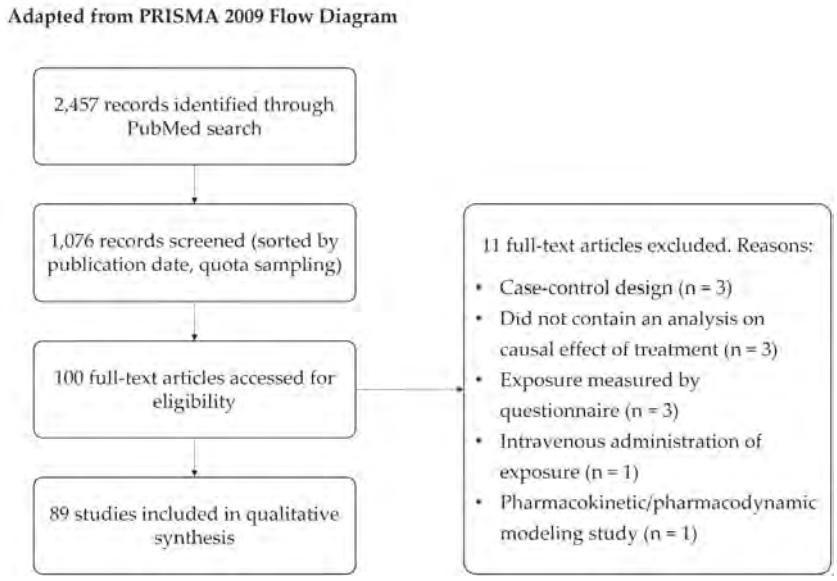


Figure 2. The screening and inclusion of eligible articles.

4 | Results

After screening the full texts of the 100 articles included during abstract and title screening, 89 studies remained based on the eligibility criteria (see Figure 2). The characteristics of the 89 included studies are summarized in Table 1. The most common patient domains considered were cardiology (17 %), neurology (11%) and primary care (10%). The median sample size was 7,011 (range 14 - 3,351,674). In 10% of studies (n = 9), a sample size calculation was reported. The length of follow-up ranged from 1 hour follow-up in one study to a median follow-up of 13.6 years in another study. Rater agreement is presented in Figure 3. Item kappas indicated that agreement between raters was low (range 0.05–0.75), which was mostly due to ambiguous reporting of the extracted information. Despite the low rater agreement of the initial scores, the presented results have a meaningful interpretation since consensus was reached for all scores with initial disagreement.

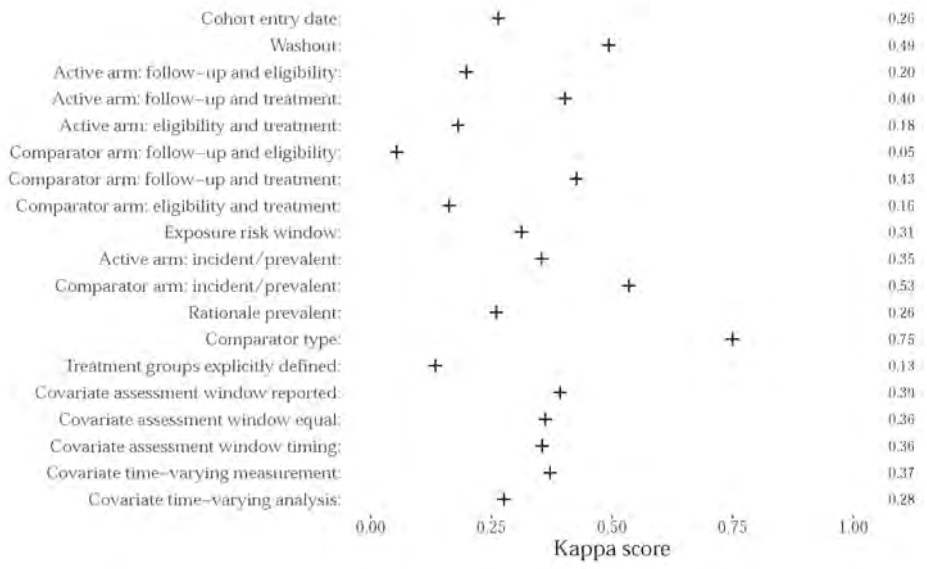


Figure 3. Agreement between raters, measured by Cohen's kappa (unweighted).

Table 1 Characteristics of the 89 included studies.

Item	Item options	Number of studies (proportion)
Journal	Annals of Pharmacotherapy	16 (0 .18)
	British Journal of Clinical Pharmacology	12 (0 .13)
	Drug Safety	11 (0 .12)
	European Journal of Clinical Pharmacology	8 (0 .09)
	Pharmacoepidemiology and Drug Safety	27 (0 .30)
	Pharmacotherapy	15 (0 .17)
Continent	Africa	1 (0.01)
	Asia	16 (0.18)
	Europe	30 (0.34)
	North America	37 (0.42)
	Oceania	2 (0.02)
	Multiple	1 (0.01)
	Not reported	2 (0.02)
Year of publication	2018	56 (0.63)
	2019	33 (0.27)
Funding	Non-pharmaceutical	83 (0.93)
	Pharmaceutical	6 (0.07)
Data source type	Claims	32 (0.36)
	Dispensing	19 (0.21)
	Hospital data	26 (0.29)
	Prescription	11 (0.12)
	Dispensing and prescription	1 (0.01)
Domain	Cardiology	15 (0.17)
	Neurology	10 (0.11)
	Primary care	9 (0.10)
	Infectious disease	6 (0.07)
	Nephrology	6 (0.07)
	Other	43 (0.48)
Sample size	< 500	23 (0.26)
	500 – 50,000	44 (0.49)
	> 50,000	22 (0.25)
Sample size calculation	No	80 (0.90)
	Yes	9 (0.10)
If sample size calculation reported, size reached?	No	1 (0.11)
	Yes	7 (0.78)
	Unclear	1 (0.11)
Cohort entry[10]	Event-based	22 (0.25)
	Exposure-based	28 (0.31)
	Multiple event-based	33 (0.37)
	Time-based	6 (0.07)
Study entry level [3, item C2]	Episode	6 (0.07)
	Person	83 (0.93)

4.1 | New-user and prevalent-user designs

An overview of item scores is given in Table 2. Forty percent of studies (95% CI 30% - 51%, n = 36) reported implementing a new-user design for both the study exposure arm and the comparator exposure arm, while 13% (7% - 22%, n = 12) reported implementing a prevalent-user design for both exposure arms (Figure 4). In 58% (42% - 74%, n =21) of studies with a new-user design for both exposure arms a washout for exposure was reported. For 6% of studies (1% - 10%, n = 5) it was unclear whether a new-user or a prevalent-user design was implemented. When a prevalent-user design was reported to be implemented, three studies provided a rationale for including prevalent users. The motivation to include prevalent users concerned biological plausibility of a cumulative effect on outcome risk[31-33].

4.2 | Alignment in new-user designs

In the 36 studies that reported implementing a new-user design in both exposure arms, moment of meeting eligibility criteria, treatment initiation, and start of follow-up were reported to be aligned in both exposure arms in 53% of studies (36% - 69%, n = 19). Moment of meeting eligibility criteria, start of treatment, and start of follow-up were reported to be misaligned in both exposure arms in 6% of studies (0% - 13%, n = 2) and alignment was unclear in 6% of studies (0% - 13%, n = 2) (Figure 3). In the remaining studies (n = 13), at least one of the six alignment items was misaligned or unclear (see Table 3 for the alignment items).

Implications of misalignment of eligibility, treatment initiation, and start of follow-up can only be assessed relative to the specified target causal effect. Initially, the protocol of this study contained an item to extract whether the target causal effect was reported, but we adjusted this during the pilot phase of our extraction tool when we discovered that no study explicitly reported an estimand (see protocol revision[23] from version 2 to version 3). Based on recommendations by the editor and reviewers, we scored whether an explicit description of the target causal effect was provided in the 36 new-user active-comparator studies. Twenty-two percent of studies (9% - 36%, n = 8) provided an explicit definition of the target causal effect. In studies that did not explicitly report the target causal effect, it was often unclear which treatment strategies were compared and which treatment decision could be informed based on evidence from the conducted study.

Table 2 Summary of reporting of information extracted from 89 reviewed articles.

Item	Item options	Number of studies	Proportion (95% confidence interval)
Study exposure arm			
New/prevalent users	New users	66	0.74 (0.65; 0.83)
	Prevalent users	14	0.16 (0.08; 0.23)
	Unclear	9	0.10 (0.04; 0.16)
Comparator exposure arm			
Comparator type	Active comparator	46	0.52 (0.41; 0.62)
	Unexposed – no use	30	0.34 (0.24; 0.44)
	Unexposed – past use	3	0.03 (0.00; 0.07)
	Combination	1	0.01 (0.00; 0.03)
	Other	6	0.07 (0.02; 0.12)
	No comparator specified	3	0.03 (0.00; 0.07)
New/prevalent users	New users	38	0.43 (0.32; 0.53)
	Prevalent users	38	0.43 (0.32; 0.53)
	Unclear	5	0.06 (0.01; 0.10)
	No comparator or symmetry design	8	0.09 (0.03; 0.15)
General design features			
Treatment groups explicitly defined	Yes	84	0.94 (0.90; 0.99)
	No	5	0.06 (0.01; 0.10)
Cohort entry date reported	Yes	71	0.80 (0.71; 0.88)
	No	18	0.20 (0.12; 0.29)
Washout reported	Yes	37	0.42 (0.31; 0.52)
	No	52	0.58 (0.48; 0.69)
Exposure risk window reported	Yes	74	0.83 (0.75; 0.91)
	No	15	0.17 (0.09; 0.25)
Covariate assessment			
Covariate assessment window reported	Yes	45	0.51 (0.40; 0.61)
	No	38	0.43 (0.32; 0.53)
	Symmetry design or self-controlled	6	0.07 (0.02; 0.12)
If covariate assessment window was reported (n = 45), was the covariate assessment window equal for all covariates	Yes	20	0.44 (0.30; 0.59)
	No	24	0.53 (0.39; 0.68)
	Not reported	1	0.02 (0.00; 0.07)
If covariate assessment window was reported (n = 45), was the covariate assessment window before initiation of treatment	Yes	27	0.60 (0.46; 0.74)
	No	13	0.27 (0.14; 0.40)
	Not reported	5	0.11 (0.02; 0.20)
If exposure was time-varying (n = 18), were covariates measured time-varying	Yes	9	0.50 (0.27; 0.73)
	No	6	0.33 (0.12; 0.55)
	Not reported	3	0.16 (0.00; 0.34)
If covariates were measured time-varying (n = 12), was this incorporated in analysis	Yes	7	0.58 (0.30; 0.86)
	No	1	0.08 (0.00; 0.24)
	Not reported	4	0.33 (0.07; 0.60)

Table 3 Summary of reporting of alignment of start of follow-up, meeting eligibility criteria and treatment initiation extracted from 36 articles that implemented a new-user design in both exposure arms.

Item	Item options	Number of studies	Proportion (95% confidence interval)
Study exposure arm			
Alignment follow-up – eligibility	Yes	24	0.67 (0.51; 0.82)
	No	9	0.25 (0.11; 0.39)
	Unclear	3	0.08 (0.00; 0.17)
Alignment follow-up – treatment	Yes	26	0.72 (0.58; 0.87)
	No	3	0.08 (0.00; 0.17)
	Unclear	7	0.19 (0.07; 0.32)
Alignment eligibility – treatment	Yes	22	0.61 (0.45; 0.77)
	No	9	0.25 (0.11; 0.39)
	Unclear	5	0.14 (0.03; 0.25)
Comparator exposure arm			
Alignment follow-up – eligibility	Yes	21	0.58 (0.42; 0.74)
	No	11	0.31 (0.16; 0.46)
	Unclear	4	0.11 (0.01; 0.21)
Alignment follow-up – treatment	Yes	25	0.69 (0.54; 0.84)
	No	5	0.14 (0.03; 0.25)
	Unclear	6	0.17 (0.04; 0.29)
Alignment eligibility – treatment	Yes	20	0.56 (0.39; 0.72)
	No	12	0.33 (0.18; 0.49)
	Unclear	4	0.11 (0.01; 0.21)

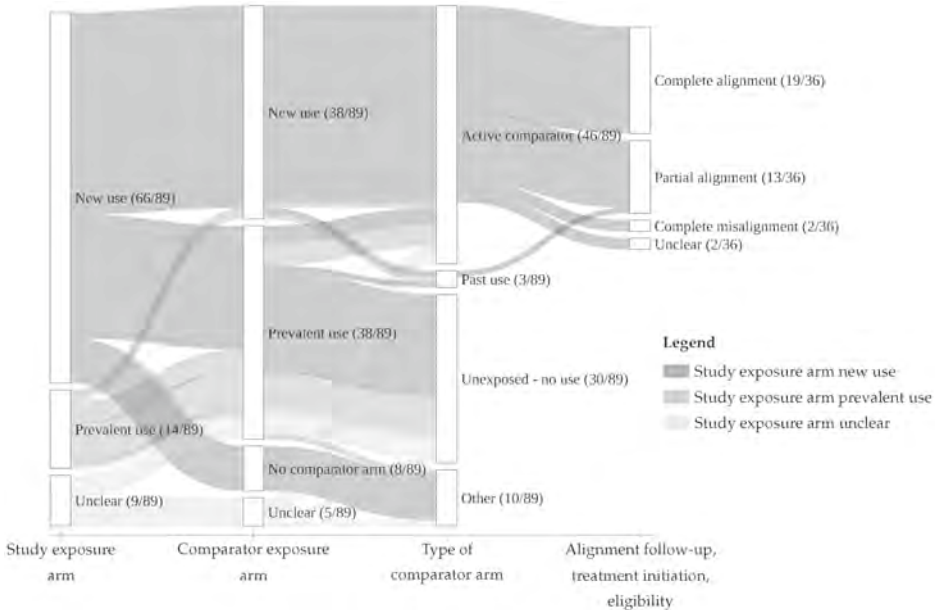


Figure 4. Frequency of reporting of implementation of new-user and prevalent-user design and type of comparator across the 89 included studies. For studies that reported implementing a new-user design, alignment of eligibility, treatment initiation and follow-up was scored 'completely aligned' when all three elements were reported to be aligned in both the active and comparator exposure arm; 'completely misaligned' when none of the elements were reported to be aligned in both the active and comparator exposure arm; 'unclear' when all three elements were unclear in both the active and comparator exposure arm; 'partial alignment' otherwise.

4.3 | Examples of good practice

Using examples from the 89 included studies, the next section illustrates how study designs that deviate from an archetypical pharmacoepidemiologic active-comparator new-user design could still provide estimates of the target treatment effect with a meaningful interpretation. We did not find any examples with a meaningfully defined study time origin among studies that contained a prevalent-user active-comparator arm.

Study design with non-user comparator arm

Korol and colleagues investigated whether initiation of spironolactone affected the risk of new onset diabetes in older patients with heart failure compared to not initiating spironolactone[34]. The patient cohort was defined by day of discharge of the first hospitalization for heart failure. The follow-up was started at the date of first dispensed prescription of spironolactone for the study exposure arm. The start of follow-up for unexposed comparator patients was inherited from the cohort entry date of the comparator and set to the time since hospital discharge from their matched comparator to establish a meaningful study time origin for non-users, given additional implementations to meet assumptions such as measuring sufficient confounders to invoke the exchangeability assumption (Table 4). Note that when an event-based cohort is established, resetting the start of follow-up at the moment of treatment initiation or comparable duration since diagnosis is essential to prevent introduction of immortal time bias[11].

Study design that anticipated immortal time

Chaignot and colleagues studied whether initiation of baclofen affected the risk of hospitalization and death compared to initiation of acamprosate in adults with an alcohol use disorder without comorbidities[35]. The patient cohort was defined by initiation of baclofen/acamprosate. To be eligible, patients had to receive at least two reimbursements for the same drug within 60 days after the first reimbursement, meaning that for included individuals, hospitalization/death could not have occurred before the second reimbursement was received. The start of follow-up was reset after the second prescription to prevent immortal time bias (Table 4). Note that the target causal effect changes by resetting start of follow-up. The study aims to identify the causal effect of baclofen compared to acamprostate given that everyone filled at least 2 prescriptions within 60 days and death was prevented during the time until they filled

a 2nd prescription. This interpretation is arguably more difficult to translate to clinical practice than a causal effect of initiating baclofen versus initiating acamprostate.

Study design that addressed time lags in start of follow-up

Belleudi and colleagues investigated whether switching from epoetin alpha (ESA α) to any other epoetin, compared to not switching, affected the risk of a blood transfusion or developing anaemia in chronic kidney disease patients[36]. The patient cohort was defined by initiation of ESA α. The follow-up was started at date of switching for the study exposure arm. A matched cohort was created to compare the risk of study outcomes in switchers versus non-switchers. The start of follow-up for non-switchers was matched to duration of ESA α treatment (± 30 days), thereby preventing time-lag bias (Table 4).

5 | Discussion

In our review of 89 pharmacoepidemiologic cohort studies of drug-outcome associations, 40% reported implementing a new-user design for both the study exposure arm and the comparator exposure arm, while 13% reported implementing a prevalent-user design in both arms, and 3 studies provided a rationale for including prevalent users. In studies that reported implementing a new-user design, we found there is room for improving alignment of meeting eligibility, treatment initiation, and start of follow-up, and reporting thereof.

It is not straightforward to understand the implications of misalignment of eligibility, treatment initiation, and start of follow-up in studies implementing a new-user design. Misalignment in the operationalization of the time origin in a study design can introduce immortal time bias or time-lag bias[3,5-7,11,12,17-19], but analytic methods can also help prevent these biases (e.g., analyzing treatment as a time-dependent variable as proposed by Suissa and Azoulay[13]). The validity of the chosen design and analysis is ideally assessed relative to the target causal effect. Since target causal effects were not often explicitly reported, we were not able to further assess implications of misalignment in the study time origin. It might have been possible to derive the target causal effect from information in the methods section in some studies. However, this would not contribute to assessment of the validity of the chosen design and analysis since target and operationalization would then overlap completely because of the reflexive definition of the target. When a target causal effect is not reported explicitly,

Table 4 Examples of design solutions for study time origin.

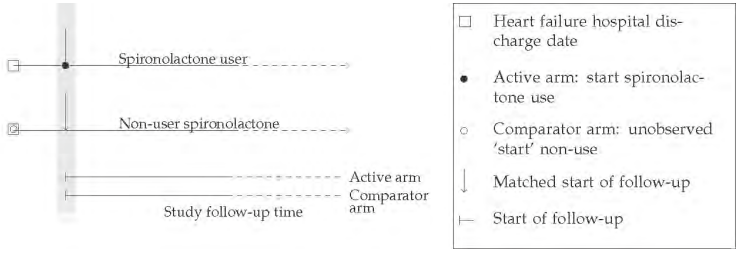
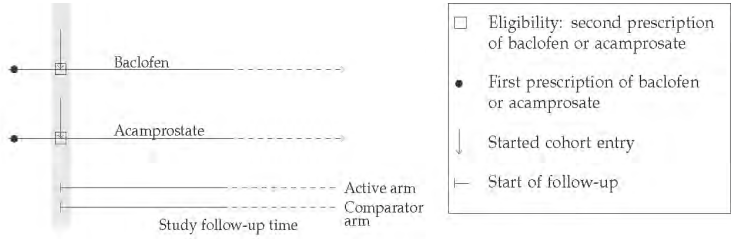
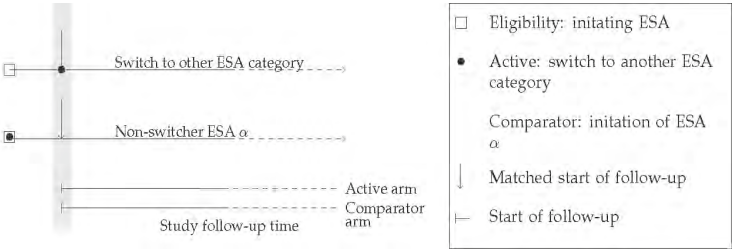
Research question	Designed time origin
Does initiation of spironolactone affect the risk of new-onset diabetes in older patients with heart failure compared to non-use of spironolactone?[34]	The patient cohort was defined by day of discharge of the first hospitalization for heart failure. For the study exposure arm, the follow-up was started at the date of first out-of-hospital dispensed prescription of spironolactone. The date of start of follow-up for unexposed comparator patients was matched to that of exposed patients on the time since hospital discharge axis to establish a meaningful study time origin for non-users. The authors did not report whether the non-user cohort was defined based on current exposure information or on future exposure information, i.e., whether non-users could still start spironolactone after their inherited date of start of follow-up or had to be unexposed during the entire study follow-up. The latter could result in a comparator cohort that is restricted to individuals who never had an indication for the treatment, which does not necessarily match the causal contrast of interest[37]. 

Table 4. *(continued)*

Research question	Designed time origin	
Does initiation of baclofen affect the risk of hospitalization and death compared to initiation of acamprosate in adults with an alcohol use disorder without comorbidities?[35]	The patient cohort was defined by initiation of baclofen/acamprosate. To be eligible, patients had to have received at least a second reimbursement for the same drug within 60 days after the first reimbursement. The start of follow-up was reset after the second prescription to prevent immortal time bias. The study thus estimates the causal effect of baclofen compared to acamprostate given that everyone filled at least 2 prescriptions within 60 days and death was prevented in the time until they filled a 2nd prescription.	
Does switching from epoetin alpha (ESA α) to any other epoetin, compared to not switching, affect the risk of a blood transfusion or developing anaemia in chronic kidney disease patients?[36]	The patient cohort was defined by initiation of ESA α. The follow-up was started at date of switching for the study exposure arm. A matched cohort was created to compare the risk of study outcomes in switchers versus non-switchers. The start of follow-up for non-switchers was matched to duration of ESA α treatment (± 30 days), thereby preventing time-lag bias (in matching, other covariates were considered as well).	

it is unclear which treatment effect the study aims to estimate, making it impossible to assess the impact of misalignment of eligibility, treatment initiation, and start of follow-up on validity of the study based on what is reported in the article. On the other hand, providing a concise and explicit definition of a target causal effect is a challenging task.

Our findings are in line with previous studies that investigated the implementation of the new-user design in specific patient domains. Yoshida and colleagues reviewed cohort studies investigating the association between use of disease-modifying antirheumatic drugs and either risk of infections (52 studies) or risk of cancers (15 studies) published between 2005 - 2015[15]. Forty percent of the studies on infection risk and 27% of the studies on cancer risk implemented a new-user active- comparator design, which is similar and lower, respectively, compared to the proportions found in our study, which covered a wider range of research areas. Suissa and Azoulay presented examples of observational studies investigating the association between metformin and cancer that suffered from immortal time bias, time-lag bias, or time-window bias[13]. Time-window bias can be an issue in case-control analysis and was not addressed here, because we only included cohort studies.

Based on our observations, it is our view that choosing a meaningful time origin is a more fundamental component of the study design than the distinction between new or prevalent users alone. Even when a new-user design was implemented, some of the articles we reviewed defined the study origin ambiguously. Reporting guidelines, such as RECORD-PE[38], state that study entry criteria and the order in which these criteria were applied to identify the study population should be clearly described. Indicating that a new-user design was implemented is insufficient to justify validity of a study design and time origin.

Our study had limitations. We focused on study-design approaches to define a meaningful study time origin. Although data analysis approaches can establish correct allocation of follow-up time as well[29,39], we did not assess them in our review. Misalignment of eligibility, treatment initiation, and start of follow-up may be appropriate when exposures are evaluated in a time-dependent manner. Four of the studies that reported implementing a new-user design studied a time-dependent exposure, thereby possibly adjusting for any misalignment in the study design. In our review, we assessed how frequently new-user and prevalent-user designs were implemented based on the reporting in original articles. It was not always possible to distinguish between lack of reporting and lack of implementation. Our results should therefore be interpreted as a summary of reporting practices on study time origin in

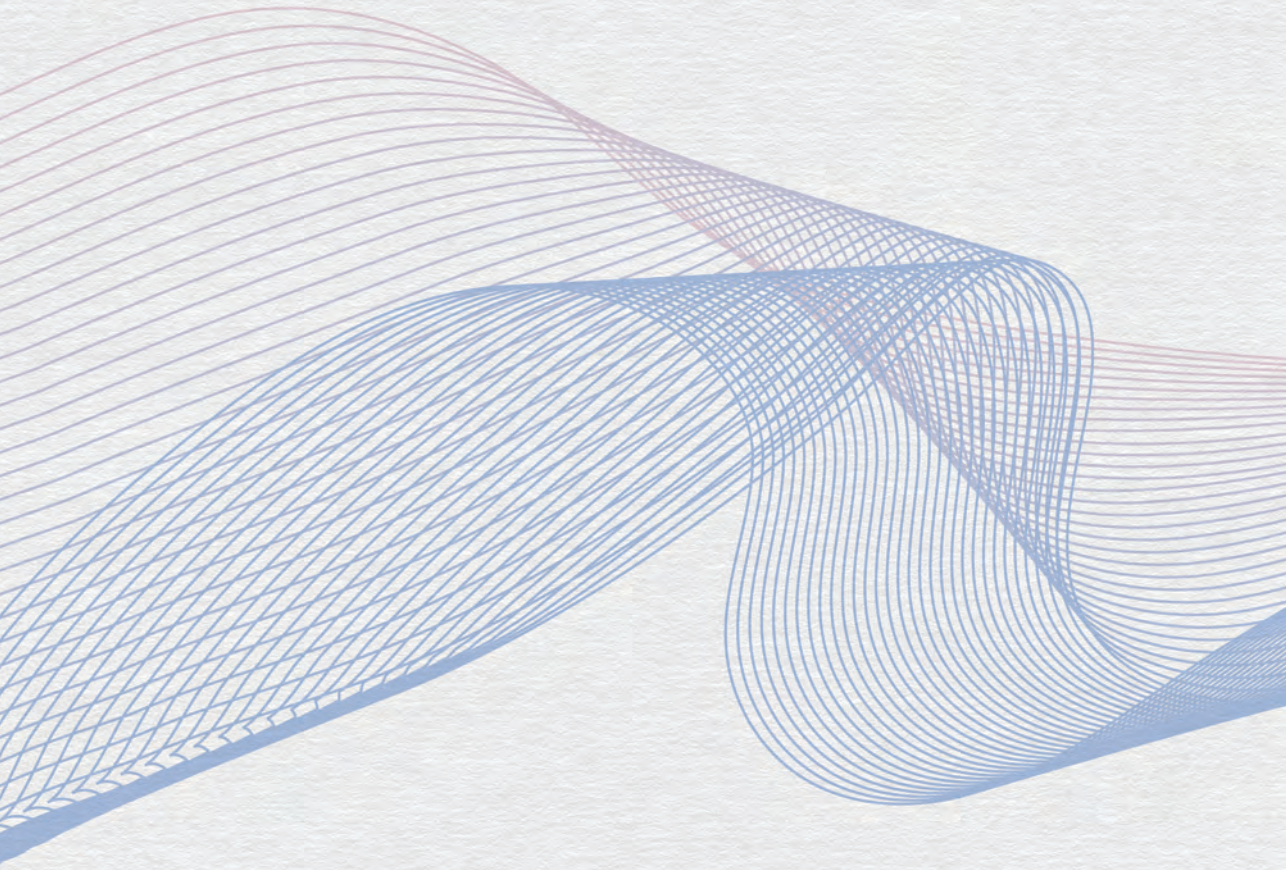
six journals. A final limitation is that our search was restricted to a convenience sample of six journals. Arguably, the six selected journals are representing the higher impact, specialist pharmacoepidemiology journals and results may therefore overestimate the quality of reporting of pharmacoepidemiologic studies in general.

The following recommendations for the design of pharmacoepidemiologic studies follow from our work. Reporting the motivation for a chosen study design and providing information on the extent to which moment of meeting eligibility criteria, treatment initiation, and start of follow-up are aligned improves the transparency and validity of research. We re-emphasize the importance of the recommendation by Schneeweiss and colleagues[40] to provide a design diagram, depicting a study's key temporal anchors and their relation to each other. When the target causal effect is unknown, it is difficult to assess whether study design and analysis are suitable for providing a meaningful estimate of the treatment effect of interest, in particular for time-dependent exposures. We recommend to explicitly report the causal contrast that is targeted in a separate statement at the beginning of the methods section. The definition of the target causal effect ideally concisely states the target population, the treatment strategies that are compared and how they are contrasted, and the outcome assessment (what and when). The causal contrast then explicates which effect is of interest (for example, an intention-to-treat effect, a per-protocol effect, an effect of treatment duration, or a comparison of treatment regimens)[26]. It should be unambiguous from this statement which future treatment decision can be informed by the study findings. Only when this information is clearly reported, the agreement can be assessed between target causal effect and applied study design and data analysis.

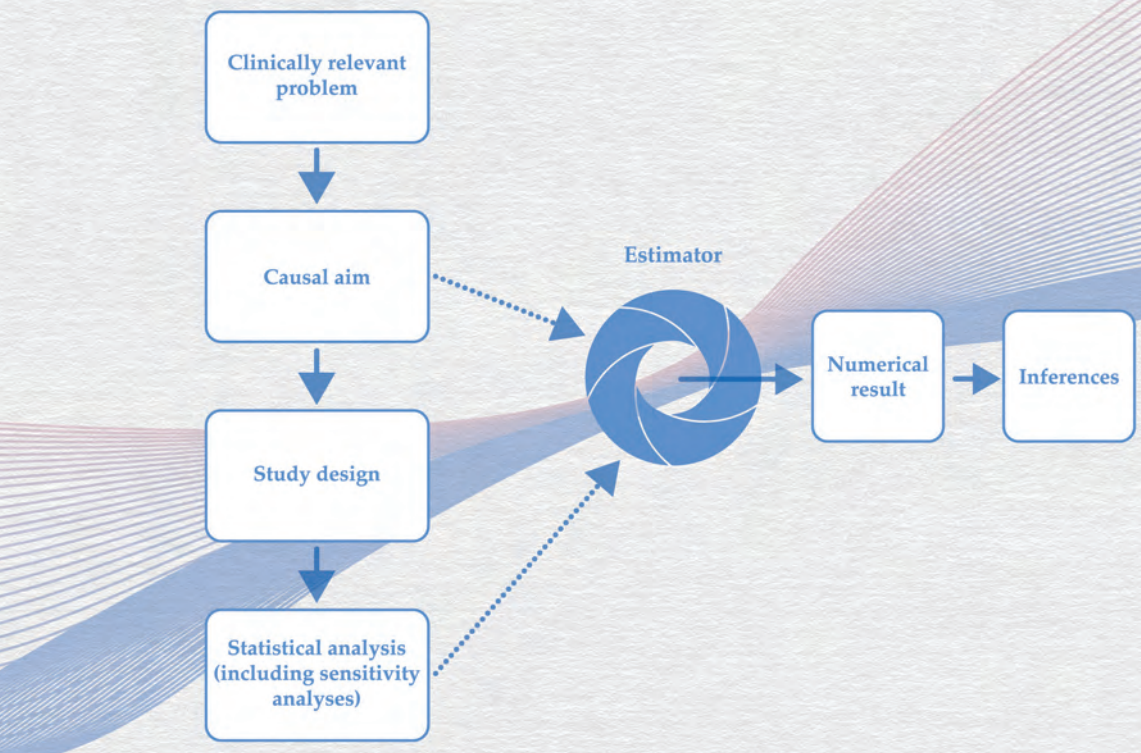
References

1. Johnson ES, Bartman BA, Briesacher BA, et al. The incident user design in comparative effectiveness research. Effective Health Care Program Research Report No. 32. (Prepared under Contract No. HHSA290200500161). AHRQ Publication No. 11(12)-EHC054-EF. *Rockville, MD: Agency for Healthcare Research and Quality.* 2012.
2. Yang W, Zilov A, Soewondo P, Bech OM, Sekkal F, Home PD. Observational studies: going beyond the boundaries of randomized controlled trials. *Diabetes Research and Clinical Practice.* 2010;88:S3-S9.
3. Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. *Current Epidemiology Reports.* 2015;2(4):221-228.
4. Wang SV, Schneeweiss S, Berger ML, et al. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1. 0. *Value in Health.* 2017;20(8):1009-1022.
5. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *American Journal of Epidemiology.* 2003;158(9):915-920.
6. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology.* 2016;79:70-75.
7. Johnson ES, Bartman BA, Briesacher BA, et al. The incident user design in comparative effectiveness research. *Pharmacoepidemiology and Drug Safety.* 2013;22(1):1-6.
8. Cox E, Martin BC, Van Staa T, Garbe E, Siebert U, Johnson ML. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report—Part II. *Value in Health.* 2009;12(8):1053-1061.
9. Roberts AW, Dusetzina SB, Farley JF. Revisiting the washout period in the incident user study design: why 6–12 months may not be sufficient. *Journal of Comparative Effectiveness Research.* 2015;4(1):27-35.
10. Vandenbroucke J, Pearce N. Point: incident exposures, prevalent exposures, and causal inference: does limiting studies to persons who are followed from first exposure onward damage epidemiology? *American Journal of Epidemiology.* 2015;182(10):826-833.
11. Suissa S. Immortal time bias in pharmacoepidemiology. *American Journal of Epidemiology.* 2008;167(4):492-499.
12. Platt R, Hutcheon J, Suissa S. Immortal time bias in epidemiology. *Current Epidemiology Reports.* 2019;6(1):23-27.
13. Suissa S, Azoulay L. Metformin and the risk of cancer: time-related biases in observational studies. *Diabetes Care.* 2012;35(12):2665-2673.
14. Hempenius M, Luijken K, de Boer A, Klungel O, Groenwold R, Gardarsdottir H. Quality of reporting of drug exposure in pharmacoepidemiological studies. *Pharmacoepidemiology and Drug Safety.* 2020;29(9):1141-1150.
15. Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. *Nature Reviews Rheumatology.* 2015;11(7):437-441.
16. Perrio M, Waller PC, Shakir SA. An analysis of the exclusion criteria used in observational pharmacoepidemiological studies. *Pharmacoepidemiology and Drug Safety.* 2007;16(3):329-336.
17. Maringe C, Benitez Majano S, Exarchakou A, et al. Reflection on modern methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. *International Journal of Epidemiology.* 2020;49(5):1719-1729.
18. Edwards JK, Hester LL, Gokhale M, Lesko CR. Methodologic issues when estimating risks in pharmacoepidemiology. *Current Epidemiology Reports.* 2016;3(4):285-296.
19. Farewell V, Cox D. A note on multiple time scales in life testing. *Journal of the Royal Statistical Society: Series C (Applied Statistics).* 1979;28(1):73-75.

20. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM. *Developing a protocol for observational comparative effectiveness research: a user's guide*. 2013.
21. Hernán MA. Counterpoint: epidemiology to guide decision-making: moving away from practice-free research. *American Journal of Epidemiology*. 2015;182(10):834-839.
22. Bosco JL, Silliman RA, Thwin SS, et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *Journal of Clinical Epidemiology*. 2010;63(1):64-74.
23. Luijken K, Spekreijse JJ, van Smeden M, Gardarsdottir H, Groenwold RHH. The use of incident and prevalent-user designs in pharmacoepidemiology: a systematic review of the literature. 2020. osf.io/wn5ad.
24. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*. 2009;6(7):e1000097.
25. Moser CA. Quota sampling. *Journal of the Royal Statistical Society Series A (General)*. 1952;115(3):411-423.
26. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*. 2016;183(8):758-764.
27. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *British Medical Journal*. 2016;355.
28. Brookhart MA. Counterpoint: the treatment decision design. *American Journal of Epidemiology*. 2015;182(10):840-845.
29. Suissa S, Moodie EE, Dell'Aniello S. Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores. *Pharmacoepidemiology and Drug Safety*. 2017;26(4):459-468.
30. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960;20(1):37-46.
31. Campbell NL, Lane KA, Gao S, Boustani MA, Unverzagt F. Anticholinergics influence transition from normal cognition to mild cognitive impairment in older adults in primary care. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*. 2018;38(5):511-519.
32. Harding BN, Weiss NS, Walker RL, Larson EB, Dublin S. Proton pump inhibitor use and the risk of fractures among an older adult cohort. *Pharmacoepidemiology and Drug Safety*. 2018;27(6):596-603.
33. Young JC, Lund JL, Dasgupta N, Jonsson Funk M. Opioid tolerance and clinically recognized opioid poisoning among patients prescribed extended-release long-acting opioids. *Pharmacoepidemiology and Drug Safety*. 2019;28(1):39-47.
34. Korol S, White M, O'Meara E, et al. Is there a potential association between spironolactone and the risk of new-onset diabetes in a cohort of older patients with heart failure? *European Journal of Clinical Pharmacology*. 2019;75(6):837-847.
35. Chaignot C, Zureik M, Rey G, Dray-Spira R, Coste J, Weill A. Risk of hospitalisation and death related to baclofen for alcohol use disorders: Comparison with nalmefene, acamprosate, and naltrexone in a cohort study of 165 334 patients between 2009 and 2015 in France. *Pharmacoepidemiology and Drug Safety*. 2018;27(11):1239-1248.
36. Belleudi V, Trotta F, Addis A, et al. Effectiveness and safety of switching originator and biosimilar epoetins in patients with chronic kidney disease in a large-scale Italian cohort study. *Drug Safety*. 2019;42(12):1437-1447.
37. Lund JL, Horváth-Puhó E, Szépligeti SK, et al. Conditioning on future exposure to define study cohorts can induce bias: the case of low-dose acetylsalicylic acid and risk of major bleeding. *Clinical Epidemiology*. 2017;9:611.
38. Langan SM, Schmidt SA, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *British Medical Journal*. 2018;363.
39. Rachet B, Abrahamowicz M, Sasco A, Siemiatycki J. Estimating the distribution of lag in the effect of short-term exposures and interventions: adaptation of a non-parametric regression spline model. *Statistics in Medicine*. 2003;22(14):2335-2363.
40. Schneeweiss S, Rassen JA, Brown JS, et al. Graphical depiction of longitudinal study designs in health care databases. *Annals of Internal Medicine*. 2019;170(6):398-406.



3


Clinically relevant problem
Causal aim
Study design
Statistical analysis (including sensitivity analyses)
Estimator
Numerical result
Inferences

What harm is there in exploration? How to distinguish pernicious ad hoc analyses from valuable scientific contributions

Exploratory analyses run the risk of being sub-optimally conducted. Since exploratory analyses are typically done aiming to generate new hypotheses, it is tempting to quickly perform a statistical test (or multiple tests) to get a first answer to the problem. However, when such 'quick-test' results are presented in a publication, their interpretation may be ad hoc and unintentionally overconfident. We provide practical pointers for good practice in exploratory etiological research, such as the use of rigorous methodologic and statistical approaches and taking responsibility for exploratory findings by reporting a clear agenda for future research.

This chapter was based on: Luijken K, Dekkers OM, Rosendaal FR, Groenwold RHH. Exploratory analyses in etiologic research: considerations for assessment of credibility. *BMJ* (in press).

1 | Background

Findings from medical research can sometimes find their way to practice very rapidly. This became clear during the outbreak of severe acute respiratory syndrome coronavirus 2, when clinical decisions sometimes had to be made on preliminary evidence combined with considerations regarding the pathophysiology of the disease. However, preliminary or exploratory findings may turn out to be incorrect and may even harm patients when implemented too early. The Hippocrates' oath "Primum non nocere" ("first, do not harm") applies to medical research just as well as it applies to clinical practice. Researchers bear responsibility for the impact their findings may have beyond the scientific debate, irrespective of the type of analyses, even if these are named 'exploratory'.

It is not uncommon to present multiple exploratory analyses in etiologic studies, generally with the aim to generate hypotheses for future research. Such hypotheses may often be considered scientifically harmless. However, even when researchers consider their study to be exploratory, a hypothesis is easily promoted to a fact. For instance, findings in journal articles can be exaggerated to more certain statements in press releases and news articles[1].

In the present paper, we discuss issues that complicate the interpretation of exploratory analyses in etiologic studies and argue that exploratory results may harm both clinical research and clinical practice. At the same time, we are aware that without exploration, there is certainly less progress in science. We provide practical pointers for researchers on how to conduct exploratory analyses and how to clarify what the exploratory results imply for future research and implementation in practice. We end with some thoughts on the delicate balance between what is pernicious and what is valuable when it comes to exploratory analyses.

2 | Exploratory analyses in etiologic research

The origin of exploratory data analysis can be traced back at least to Tukey in the 60's and 70's[2,3], who encouraged statisticians to develop visualization techniques for representing and capturing structures in data sets to establish new research questions. Tukey pioneered in motivating the value of data-driven questions and the development of methods for improving non-specific knowledge about these questions to more exact answers. In this, he seemed to be predominantly concerned with science more than

decision-making. His writing paid little attention to the role of complex models and increasing computing power – two aspects that allowed for more extensive exploratory research over the course of time.

In what follows, we use the term *exploratory analyses* to indicate analyses that provide preliminary information that will help defining new research questions, and which are not always specified prior to data analysis. Exploratory analyses are often conducted additional to planned primary analyses of a study, which we denote *confirmatory analyses*. Sensitivity analyses, in which the main hypothesis is evaluated under different assumptions, are not considered to be exploratory in this paper. Outcomes that are evaluated as a secondary objective but are correlated to the primary outcome are not considered exploratory either, because these analyses contribute to the investigation of the primary research question. Genome-wide association studies, in which the exploratory nature of is commonly accounted for by addressing multiple testing[4], are beyond the scope of this paper.

For randomized trials, preregistration of the study protocol is considered the norm[5] and guidance on cautious interpretation of subgroup analyses is increasingly available[6] as is guidance for reporting of exploratory work preceding the randomized trial in a feasibility or pilot study[7]. However, in observational etiologic research, exploratory research questions may arise during data-exploration or statistical analysis, and little guidance exists on how these questions should be studied and interpreted.

Published etiologic studies often contain numerous results. As an illustration, in the first issue of 2021 from four major epidemiology journals (25 original etiological articles from the American Journal of Epidemiology, Epidemiology, European Journal of Epidemiology and International Journal of Epidemiology), we found that these articles presented on average 33 (range 1 – 120) associations for the primary analysis, on average 30 (range 0 – 336) associations for sensitivity analyses, and on average 163 (range 0 – 1467) associations in additional analyses, mainly concerning subgroup or interaction analyses (details in Online Supplement). Categorizing the additional analyses as either 'confirmatory' or 'exploratory' was not straightforward. Most articles did not explicitly report which analyses were prespecified and only one study referred to a publicly available protocol[8]. Some subgroup analyses seemed to have been carried out thoughtfully, with the intention to evaluate exposure effect heterogeneity among well-established subgroups, while other subgroup analyses seemed to have been performed exhaustively across a large number of possible risk factors, and results were selectively reported in the main text based on statistical significance.

As the distinction confirmatory / exploratory is not always straightforward in etiologic research, we propose to describe analyses in terms of a continuum of scrutiny (Figure 1), relating to exploratory and confirmatory studies alike. Where an analysis is situated on this spectrum from 'ad hoc' to 'targeted' depends on the nature of the research question and methodological rigor of the analysis. An analysis qualifies as being 'targeted' when the research question is well-advised by theory and the methodology and statistical analysis are designed accordingly, leading to fair credibility of the resulting evidence. While exploratory analyses are generally situated more on the 'ad hoc' end of this scale, they can be moved towards the 'targeted' side by conducting the study rigorously, as is described in more detail later.

3 | Exploratory analyses require directions for the reader

Exploratory analyses seem to be perceived harmless if their "hypothesis-generating" nature is made clear. No doubt, an open-minded approach in exploratory research leads to new insights and scientific progress that cannot be achieved by a rigid system of confirmatory research alone[9]. Even though confirmation can be considered a safeguard against the incorrect promotion of a hypothesis to a fact, there is a practical downside to such an approach as it is not feasible to conduct a confirmatory study for each of the large number of hypotheses that are currently generated in exploratory analyses. It can be easily understood that if a confirmatory study is needed for each of the aforementioned 163 results from additional analyses that we found on average in recently published epidemiological studies, this would overcharge available research capacity, if only in terms of independent data sets.

Exploratory results are tentative findings that should not form a basis for implementation in clinical practice. Some associations discovered during exploration of the data are expected to be false positive findings. Rapid implementation of exploratory findings into policy can have unwarranted consequences. Delay in dissemination of knowledge from science to clinical practice may detain patients from health benefits, but informing medical decisions based on preliminary evidence should be the exception rather than the rule to avoid harm by scientifically unjustified policies given the risk of false positive results.

Presentation of multiple analyses requires pointers for readers. For a judgement on the credibility of exploratory findings, readers of a study depend on unambiguous reporting of the nature and conduct of exploratory analyses, from which they can assess

whether the scrutiny of the analysis was 'ad hoc' or 'targeted' (Figure 1). Particularly for those that are more of the 'ad hoc' type, it is the responsibility of the author to prioritize which of the results are worth conducting a confirmatory study for and to report what that prioritization is based on (e.g., pathophysiological mechanism). Taking this responsibility clarifies how research resources should be allocated under minimally expected research waste[10-12].

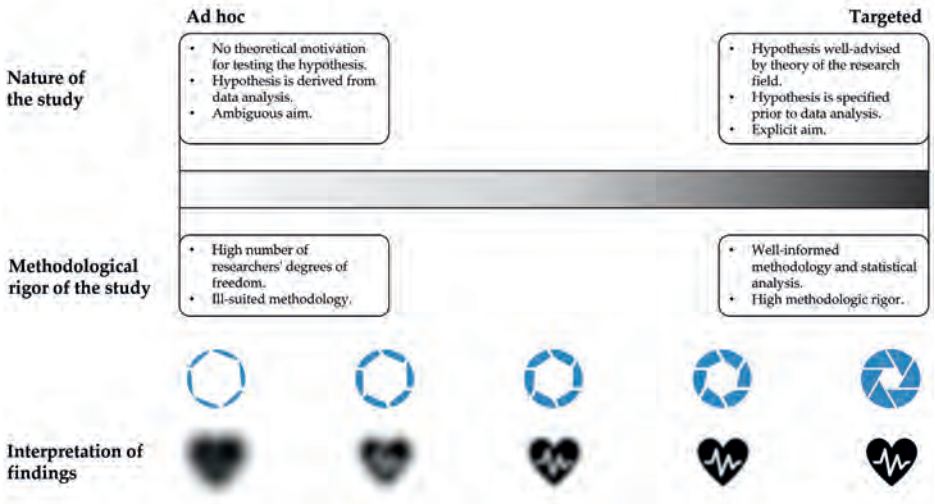


Figure 1. The continuum of scrutiny in conduct of etiologic studies. The nature of a research question and the degree to which it is aligned with the study design and statistical analysis determine the degree to which findings can be interpreted in a meaningful way. Although exploratory questions are generally situated more on the 'ad hoc' side of this scale, the interpretation of results could be clarified by improving methodological quality of the data analysis.

4 | 'Exploratory' does not imply 'less rigorous'

Prioritization of future research cannot be done based on numeric results of exploratory analyses only, because seemingly convincing results easily fool us into taking an observed association as something real and finding a clinical explanation that does not follow from the statistical evidence[13,14]. The statistical properties of exploratory tests are less well known than those of confirmatory tests[15]. For instance, the expected number of false positives (i.e., type I error rate) is likely inflated when statistical tests are not specified prior to data analysis or when the choice for a particular test depends on patterns in the data. While procedures have been developed for correction of multiple

testing in confirmatory settings, consensus on how to prevent false positive findings in exploratory settings has not been established.

What is more, the design and methods applied in an exploratory analysis may be less optimal than for the primary analysis of the study, which further complicates interpretation. For instance, when various exposure-outcome associations are explored, this likely requires more consideration than combining another set of covariates in a model. Analytic decisions should be reconsidered for each exposure-outcome combination that is studied, for instance the selection of confounders, specification of functional forms, model specification, and evaluation whether assumptions can be invoked[16,17].

The fact that an analysis is exploratory does not mean it can be taken as a loophole to avoid setting up a rigorous statistical analysis plan. Designing a study to test a hypothesis usually requires time and effort. When additional analyses are performed, either more resources should be spent to execute and report them rigorously, or it should be determined upfront how to restrict the number of analyses to match the available research capacity. Probably, to impose a strict methodological standard for exploratory analyses will induce a reduction in these analyses.

5 | Some good practices for exploratory analyses

Table 1 describes recommendations for good practice in conducting exploratory analyses at different stages of study conduct.

5.1 | Protocol

Protocols may contain a section describing exploratory analyses. It is tempting to presume that the plan to analyze a hypothesis will clarify itself once the data can be accessed. Of course, not every detail can be thought of and specified in advance, but interpretation of results provided by data can be challenging when no question was clearly articulated prior to seeing the answer. A way to prevent this confrontation is to pre-specify the analysis as thoroughly as possible[18,19]. The continuum of scrutiny suggests that pre-specification of a hypothesis (or an exposure – outcome relation) alone is insufficient for an analysis to qualify as 'targeted' and to render credible results. Furthermore, at the design stage of the study, it might be worthwhile to consider the degree of information that can be gained from exploratory analyses: if they will

provide little extra knowledge, why not refrain from performing these analyses? If truly interesting, why not work out a detailed protocol focused on that research question?[20,21]

Table 1 Practical recommendations for exploratory analyses

Research stage	Recommendation
Study protocol	- Limit the number of exploratory analyses. - Prespecify the aims and conduct of data analysis as thoroughly as possible in a research protocol. - Preregister the protocol (see [8] for an example of good practice*).
Statistical analysis	- When the idea to perform an analysis comes up after running the primary analysis, take a time-out to establish a targeted analysis plan for each exploratory research question. - Avoid conducting analyses ad hoc by formulating and analyzing exploratory questions as rigorous as possible. - Make sure that inferences are supported by the applied methodology and in line with their nature (i.e., confirmatory or more exploratory).
Reporting	- Present the results of every exploratory analysis including the methodological rigor with which the question was examined (ad hoc or targeted), possibly in supplementary files (see [23] for an example of good practice*). - Set a research agenda: which of the generated hypotheses are worth studying in confirmatory follow-up research, give arguments for that and direction on how should they be studied? (see [24] for an example of good practice*).
Peer-review and journals	- Credit methodological rigor rather than the number of results reported.

* The examples of good practice were identified from etiologic studies published in the first issues of 2021 of four major epidemiological journals (see Online Supplement).

5.2 | Statistical analysis

An obvious yet relevant good practice in performing statistical analyses is to stick to the prespecified analysis plan as closely as the data allow for[22]. In the situation where the idea to perform an analysis comes up after running the primary analysis, i.e., in case of a *post-hoc* analysis, it is of importance to design and analyze this exploratory question as rigorously as possible, similar to the way a confirmatory analysis would have been planned. We recommend taking a time-out to establish a targeted analysis plan for each exploratory research question, to avoid performing post-hoc analyses using ad hoc methodology. Importantly, the results need to be interpreted in line with their exploratory nature and communicated (e.g., in a research paper) as such.

Additional exploratory analyses introduce issues regarding multiplicity of analyses that have no straightforward solution but should be taken into consideration.

Which analyses should be statistically corrected for multiple testing and in what way? Should the planned primary analyses be corrected for multiple testing once additional exploratory analyses are performed? Performing multiple tests without a statistical correction inflates the risk of drawing false-positive conclusions, but too strict correction for multiple-testing can increase the probability of false-negative findings too (i.e., the type II error rate)[25]. This could occur, for instance, when an analysis of various positively correlated hypotheses is corrected for multiple testing as if all hypotheses were independent (for example by applying a *Bonferroni correction*). The decision to statistically correct for multiple testing depends on i.a. the total number of tests performed in the same dataset, the correlation between the hypotheses being tested, and the sample size.

5.3 | Reporting

Apart from clearly indicating which analyses were exploratory in nature, the report of a study should provide further guidance on the level of evidence for each of the hypotheses tested. A researcher is eminently aware to what degree the quality of the data and statistical methods were suitable to answer the research question and therefore can judge firsthand the credibility of exploratory finding[26]. Since the prior likelihood of an effect is generally low in exploratory observational studies, the posterior evidence is likely not astonishingly credible. Such knowledge can be communicated by stating a research agenda containing prioritization of future research and how this should be set up, thus allowing researchers to take responsibility for the presented exploratory findings and future research that should be performed, avoiding the empty statement that 'more research is needed'.

5.4 | Peer review and publication

Results are still often decisive for publication of a study[27] and presentation of multiple exploratory findings seems to match this reward system; articles that contain myriad results might contain interesting findings that make the study attractive for publication. A study that carefully specifies a sound research question and implements apt methodology may present fewer results but can make a valuable contribution and is potentially less harmful. Peer reviewers and journal editors should take this into account during the evaluation of a manuscript submitted for publication.

By no means should a manuscript reporting on a multitude of analyses be rejected for publication. Sensitivity analyses evaluating the primary analysis under different

assumptions can still lead to numerous results, but these analyses add to the credibility of the conclusions. Moreover, although allowing studies to report only a handful of additional analyses may improve the methodological quality (as there is more attention for each analysis separately), this may do more harm than good if coherent bodies of results become fragmented into multiple publications ('salami-slicing'[28]).

6 | Balancing opportunities and perils of exploratory analyses

Exploration is indispensable to the progress of science. Strict confirmatory studies are a powerful mechanism for final evaluations of existing evidence before implementation in clinical practice, yet will likely not spark new ideas[29]. In other words, research that is situated at the extreme targeted side of the continuum presented in Figure 1 provides most straightforward interpretations, but likely adds little to scientific understanding and progress. Open-minded exploratory analyses can lead to serendipitous discoveries and resourceful innovations of epidemiological science. Yet, this requires that exploratory questions are being answered using rigorous methodologic and statistical approaches. Even then, implications for clinical practice remain uncertain (and must so). Especially in medical science, where study results are sometimes quickly implemented in clinical practice, it is essential that researchers take responsibility for the results they report by minimizing the number of ad hoc analyses, designing all analyses thoroughly, and clearly explaining which exploratory findings should be investigated in future research and how. Only when exploratory analyses are conducted and interpreted *lege artis* will they unfold their full value.

Online Supplementary Files

The supplementary files referred to in this Chapter are available online at https://github.com/KLuijken/Dissertation_Online_Supplements/tree/main/Chapter_3

References

1. Sumner P, Vivian-Griffiths S, Boivin J, et al. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *British Medical Journal*. 2014;349.
2. Tukey JW. The future of data analysis. *The Annals of Mathematical Statistics*. 1962;33(1):1-67.
3. Tukey JW. *Exploratory Data Analysis*. Vol 2: Reading, Mass.; 1977.
4. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*. 2014;15(5):335-346.
5. Zarin DA, Tse T, Williams RJ, Rajakannan T. Update on trial registration 11 years after the ICMJE policy was established. *New England Journal of Medicine*. 2017;376(4):383-391.
6. Kent DM, Paulus JK, Van Klaveren D, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of Internal Medicine*. 2020;172(1):35-45.
7. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *British Medical Journal*. 2016;355.
8. Keys MT, Serra-Burriel M, Martínez-Lizaga N, et al. Population-based organized screening by faecal immunochemical testing and colorectal cancer mortality: a natural experiment. *International Journal of Epidemiology*. 2021;50(1):143-155.
9. Tukey JW. We need both exploratory and confirmatory. *The American Statistician*. 1980;34(1):23-25.
10. Schuit E, Roes KC, Mol BW, Kwee A, Moons KG, Groenwold RH. Meta-analyses triggered by previous (false-) significant findings: problems and solutions. *Systematic Reviews*. 2015;4(1):1-8.
11. Lakens D. The reproducibility project: a model of large-scale collaboration for empirical research on reproducibility. In Stodden V, Leisch F, Peng RD, editors. *Implementing reproducible research*. Osa Roca: Taylor and Francis Ltd. 2014:299-324. (Chapman & Hall/CRC The R Series).
12. Ioannidis JP. Why most published research findings are false. *PLoS medicine*. 2005;2(8):e124.
13. Goldacre B, Drysdale H, Marston C, et al. COMPare: Qualitative analysis of researchers' responses to critical correspondence on a cohort of 58 misreported trials. *Trials*. 2019;20(1):1-13.
14. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*. 2016;31(4):337-350.
15. Goeman JJ, Solari A. Multiple testing for exploratory research. *Statistical Science*. 2011;26(4):584-597.
16. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*. 2013;177(4):292-298.
17. Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I, initiative ttgCIotS. Formulating causal questions and principled statistical answers. *Statistics in Medicine*. 2020;39(30):4922-4948.
18. Shmueli G. To explain or to predict? *Statistical Science*. 2010;25(3):289-310.
19. Wagenmakers E-J, Wetzels R, Borsboom D, van der Maas HL, Kievit RA. An agenda for purely confirmatory research. *Perspectives on Psychological Science*. 2012;7(6):632-638.
20. Loder E, Groves T, MacAuley D. Registration of observational studies. British Medical Journal. 2010;340:c950.
21. Wang SV, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *British Medical Journal*. 2021;372.
22. Huebner M, le Cessie S, Schmidt CO, Vach W. A contemporary conceptual framework for initial data analysis. *Obs Stud*. 2018;4:171-192.
23. Rojas-Saunero LP, Hilal S, Murray EJ, Logan RW, Ikram MA, Swanson SA. Hypothetical blood-pressure-lowering interventions and risk of stroke and dementia. *European Journal of Epidemiology*. 2021;36(1):69-79.

24. Morrison CN, Kaufman EJ, Humphreys DK, Wiebe DJ. Firearm homicide incidence, within-state firearm laws, and interstate firearm laws in US counties. *Epidemiology.* 2020;32(1):36-45.

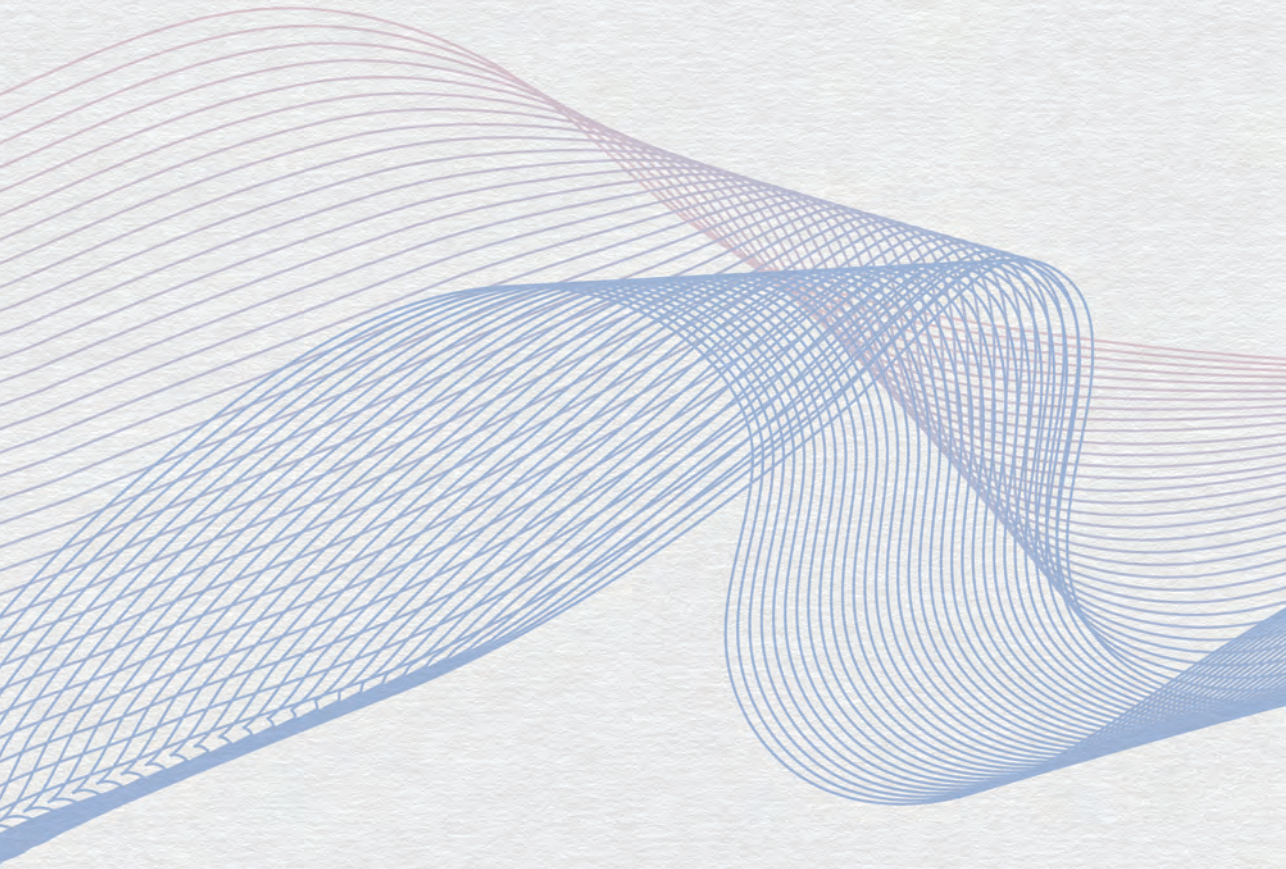
25. Groenwold RH, Goeman JJ, Le Cessie S, Dekkers OM. Multiple testing: when is many too much? *European Journal of Endocrinology.* 2021;184(3):E11-E14.

26. Mayo DG, Spanos A. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science.* 2006;57(2):323-357.

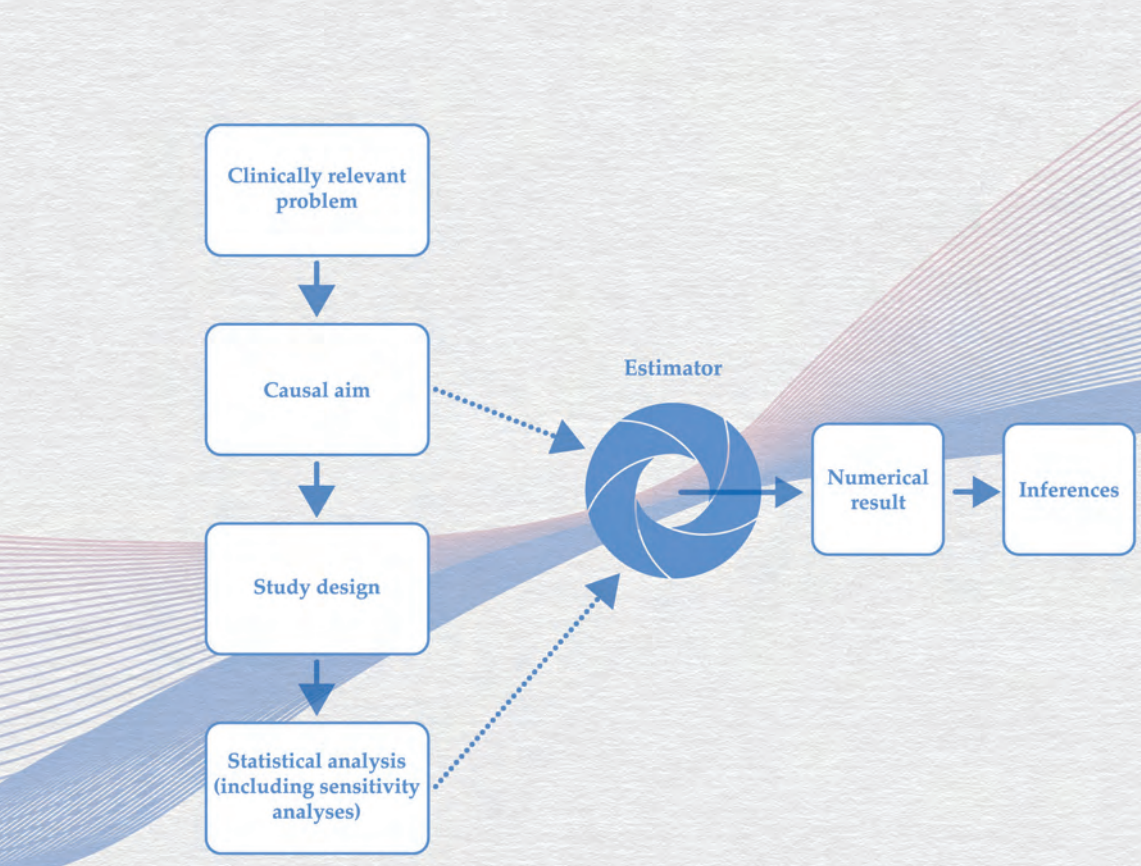
27. Smaldino PE, McElreath R. The natural selection of bad science. *Royal Society open science.* 2016;3(9):160384.

28. Editorial. The cost of salami slicing. *Nature Materials.* 2005;4(1):1.

29. Vandenbroucke JP. Observational research, randomised trials, and two views of medical science. *PLoS med*. 2008;5(3):e67.



4


Clinically relevant problem
Causal aim
Study design
Statistical analysis (including sensitivity analyses)
Estimator
Numerical result
Inferences

A comparison of full model specification and backward elimination of potential confounders when estimating marginal and conditional causal effects on binary outcomes from observational data

A common view in epidemiology is that automated confounder selection methods, such as backward elimination, should be avoided as they can lead to biased and overly precise effect estimates. Nevertheless, backward elimination remains regularly applied. We investigated if and under which conditions causal effect estimation in observational studies can improve by using backward elimination on a prespecified set of potential confounders. An expression was derived that quantifies how variable omission relates to bias and variance of effect estimators. Additionally, 3,960 scenarios were defined and investigated by simulations comparing bias and mean squared error (MSE) of the conditional log odds ratio, log(cOR), and the marginal log risk ratio, log(mRR), between full models including all prespecified covariates and backward elimination of these covariates. Applying backward elimination resulted in a mean bias of 0.03 for log(cOR) and 0.02 for log(mRR), compared to 0.56 and 0.52 for log(cOR) and log(mRR), respectively, for a model without any covariate adjustment, and no bias for the full model. In less than 3% of the scenarios considered, the MSE of the log(cOR) or log(mRR) was slightly lower (max 3%) when backward elimination was used compared to the full model. When an initial set of potential confounders can be specified based on background knowledge, there is minimal added value of backward elimination. We advise not to use it and otherwise to provide ample arguments supporting its use.

This chapter was based on: Luijken K, Groenwold RHH, van Smeden M, Strohmaier S, Heinze G. A comparison of full model specification and backward elimination of potential confounders when estimating marginal and conditional causal effects on binary outcomes from observational data. *Biometrical Journal* (in press).

1 | Background

Identification of causal effects from observational data relies on proper control for confounding. It is generally advised that confounders are determined based on the causal structure of the data, about which one may possess background knowledge, or one could at least make defendable assumptions[1-4], and that automated covariate selection methods, such as stepwise selection and backward elimination, should be avoided as they can lead to seriously biased estimated effect sizes and underestimation of statistical uncertainty by model-based confidence intervals (CIs)[5,6].

Despite these warnings, automated selection procedures for selection of confounders remain widely applied[7-13]. One reason for the popularity may be that, at least in theory, using automated selection as an add-on to selection of potential confounders based on background knowledge may lead to improved efficiency[2,14,15]. For instance, backward elimination has occasionally been reported to improve estimation in terms of mean squared error (MSE) of the effect estimator[16]. Limited guidance exists about when backward elimination could be beneficial in observational studies in which confounding adjustment is needed[17-19].

The aim of the current study was to extend recommendations for practicing statisticians on the use or avoidance of automated variable selection for descriptive models provided by Heinze and colleagues[5] to a causal inference context. Specifically, we compare the efficiency of causal effect estimation by multivariable modelling in observational studies when fitting a model with all potential confounders (full model) compared to using backward elimination (see Box 1). Out of the many available methods for variable (or confounder) selection, we focus on backward elimination, since it is widely implemented in statistical software packages and is often considered superior to alternatives such as univariable screening, forward and stepwise selection[20]. We focus on outcome-oriented selection of confounders, meaning that exposure-oriented selection procedures, for instance as part of propensity score methods, are beyond the scope of this article. Furthermore, we assume that sufficient clinical expertise is available to specify an outcome model with covariates presumably related to the exposure and/or outcome. This model is assumed to include at least all such covariates and to correctly specify all nonlinear covariate-outcome relations but may include covariates only related to the exposure (instruments) and/or true confounders, or irrelevant covariates.

In Section 2, we present comparative analyses of a motivating example. In Section 3, we discuss arguments in favor of and against the use of backward elimination as a means of automated selection among potential confounders. In Section 4, we perform simulation studies to investigate whether there is a benefit of using a backward-elimination estimator compared to a full-model estimator to estimate the target causal effect. We end with a discussion of the implications for clinical research.

Box 1. Motivation to compare backward elimination of potential confounders neutrally with a full model

- After identifying a set of potential confounders, uncertainty about the causal role of some covariates may remain. Although backward elimination can reduce the adjustment set, it assumes that the initial set of potential confounders is a sufficient adjustment set.
- The disjunctive cause criterion by VanderWeele and Shpitser can guide confounder selection[15]. This criterion states to control for covariates that are either a cause of the exposure or a cause of the outcome, which may lead to adjustment for instrumental variables. Therefore, they recommended implementing backward elimination or forward selection to eliminate such variables. On the other hand, Vansteelandt and colleagues argue that instrumental variables should not necessarily be eliminated from the adjustment set, because the uncertainty they introduce on the estimated exposure effect may reflect lack of information about the effect of interest[18].
- Greenland and colleagues proposed to compare a model adjusted for a sufficient set of confounders where one confounder is deleted by hand to a full model by estimatingthe change in mean squared error (MSE) which was illustrated in an empirical data set[14]. Since similar bias and variance considerations apply to backward elimination, it is worthwhile to compare a full model and use of backward elimination in more settings.
- Backward elimination has been reported to improve estimation in terms of MSE of the effect estimator[16].

2 | Motivating example: coronary artery bypass grafting study

We illustrate confounder selection using a study that investigated the causal effect of a computer tomography angiography (CTA) examination of the main coronary artery prior to coronary artery bypass grafting (CABG) surgery on the postoperative stroke risk of a patient[21]. We used a simulated dataset based on the empirical data (details in[21]) that was previously used for methodological work[22]. In the simulated dataset, the sample size and relationships between the variables were preserved and similar to the original dataset. In Online Supplementary File 1, we provide Rcode to allow replication of this example.

2.1 | Defining causal estimands

We defined two research questions and the corresponding estimands[23]. The first research question compared the risk of post-operative stroke for patients with known characteristics when refraining from screening for aortic disease using CTA prior to CABG surgery versus the risk when patients were screened using CTA. The causal contrast, no CTA screening versus CTA screening given a set of characteristics, can, for instance, be expressed as a conditional risk difference, a conditional risk ratio or a conditional odds ratio (cOR). We defined the estimand as the cOR.

The second question of interest concerned the effect of not exposing an entire target population to CTA screening versus exposing everyone to CTA screening. The causal contrast could be expressed as a marginal risk difference, a marginal risk ratio (mRR) or a marginal odds ratio. We defined the estimand as the mRR.

2.2 | Linking the observed data to the estimand

Whether a causal effect can be identified from observational data depends, among other things, on the extent to which confounding can be adjusted for. Heinze and colleagues recommended to generate an initial working set of covariates based on clinical expertise and background knowledge, without yet using the dataset at hand[5,24]. In studies of causal inference, it is often helpful to visualize assumed causal dependencies between covariates, where the level of formalization of those dependencies may sometimes reach that of a directed acyclic graph (DAG) (we refer to[25] for recommendations on implementation). In doing so, a researcher explicates knowledge about variables that are *irrelevant* to the study question, as leaving out variables is a stronger assumption

than including them. Accordingly, for covariates that are included in an initial working set, many decisions are still to be made regarding their causal role and relevance.

In the original study[21], the initial working set contained 23 measured covariates that described the health state of a patient just before the decision to perform CTA or not. Detailed causal assumptions that could be represented in a DAG were not supported by the cross-sectional assessment of these covariates, but we could exclude collider stratification bias or presence of mediators when using these covariates as a confounding adjustment set.

2.3 | Defining causal estimands

We estimated the cOR by the exponentiated regression coefficient of no CTA screening in a multivariable logistic regression model with Firth's correction[26-28] (CIs based on profile penalized likelihood) including the 23 covariates specified in the initial working set. We estimated the mRR based on predictions of potential outcomes from that multivariable logistic regression model[29-31] (CIs based on 500 bootstrap samples using the percentile method). Additionally, we applied data-driven selection of the 23 prespecified covariates by means of backward elimination at a significance level of 0.157 approximating selection by the Akaike information criterion[5]. For the backward-elimination estimator, we contrasted 'selected-model' CIs, which condition on the finally selected covariates, to 'global' bootstrap CI, where the selection process was repeated in each bootstrap resample. The selected-model CIs were based on profile penalized likelihood for the cOR and computed from fitting the finally selected model in 500 bootstrap samples using the percentile method for the mRR.

In this example, backward elimination reduced the adjustment set by eight potential confounders. While for both cOR and mRR the full-model CIs were wider than the (invalid) selected-model CIs, the global bootstrap CIs were the widest (Table 1). Clearly, additional variability arises from the uncertainty in the selection which must be captured by repeating the selection process in each bootstrap resample. Heinze and colleagues proposed to evaluate bias and added uncertainty by two bootstrap-based measures, relative conditional bias (RCB) and root mean squared difference ratio (RMSDR)[5,32]. In our example, the RCB for the log cOR was -1.3% and RMSDR was 1.06. The RCB for the log mRR was -2.6% and RMSDR was 1.07. These measures also indicated a possible variance inflation by using backward elimination.

Table 1 Results for the CABG study.

Estimand	Model	Estimate	Confidence interval estimation approach	95% Confidence interval	Confidence interval width (upper/lower)
cOR	Full	4.48	PPL	[2.10, 10.01]	4.77
	Selected	3.83	Invalid: selected-model PPL	[1.99, 7.77]	3.90
			Global bootstrap	[2.10, 11.25]	5.36
mRR	Full	3.68	Bootstrap	[2.02, 7.09]	3.51
	Selected	3.24	Invalid: selected-model bootstrap	[1.87, 6.19]	3.31
			Global bootstrap	[1.93, 7.17]	3.72

Abbreviations: cOR = conditional odds ratio, mRR = marginal risk ratio, PPL = profile penalized likelihood

3 | Use of automated covariate selection

3.1 | Arguments in favor of automated selection of confounders

Bias and variance of an effect estimator can be combined in a single measure; the MSE. The MSE can be interpreted as the expected value of the squared distance of an estimate to the true value, which can be alternatively expressed as $MSE = bias^2 + variance$. For a linear regression model, the value of omitting a covariate in terms of reducing the MSE of an effect estimator can be quantified directly (see Appendix). We provide a simplified representation of this principle here that extends to settings with binary outcomes.

Consider a setting with an outcome, an exposure and one covariate. The effect of the exposure on the outcome is evaluated under two estimation strategies: 'always include the covariate' (full) versus 'always omit the covariate' (omit). Assuming that the bias in the exposure effect estimator of the 'full' strategy is 0, in terms of MSE we find a benefit in omitting the covariate if, for the effect of the exposure on the outcome, the following inequality holds:

$$Bias^2_{omit} < Variance_{full} - Variance_{omit} \quad (1)$$

If (1) holds, the reduced variance of the 'omit' strategy outweighs the increase in squared bias, and thus there is a benefit of omitting the covariate in terms of MSE, and hence produces a more efficient estimate. If we ignore a possible small sample

bias[33,34], only the righthand side of (1) is inversely proportional to sample size. Thus, there should be a threshold sample size n, such that (1) holds for all values smaller than that n. Figure 1 illustrates this phenomenon. Figure A3 and A4 (in Appendix) illustrate that n increases with a stronger association between the exposure and the covariate, with a weaker association between the outcome and covariate, and with a lower variance of the exposure variable.

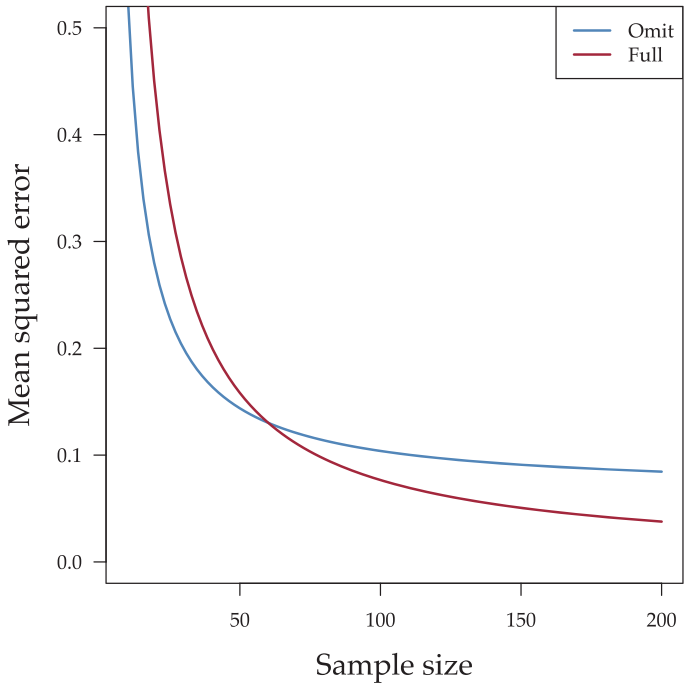


Figure 1. Illustration of the bias-variance trade-off for the ordinary least squares estimator of the exposure effect when including (Full) or omitting (Omit) covariate *L*. The blue and red line are computed using expressions for the mean squared error under the 'Omit' and 'Full' strategy, respectively (Appendix). The value of n for which the reduced variance by omitting L outweighs the increase in squared bias is around 60. This illustrates inequality (1) in the main text for a linear model and ordinary least squares estimation. For sample sizes < 60, omission of the covariate resulted in a lower mean squared error of the exposure effect estimator in a linear setting.

3.2 | Arguments against automated selection of confounders

Selection of variables by statistical procedures is sometimes incorrectly thought to be a prerequisite for model building[35]. However, a 'statistically significant' result neither confirms whether a covariate is indeed a confounder, nor does insignificance prove that it is not. A well-known counterargument against use of data-driven selection of confounders is that the causal structure of the data cannot be derived from observed

associations only. For example, a covariate has a different causal status being a *confounder* compared to being a *mediator*, but in both cases, it may be statistically associated with the exposure and/or outcome. Automated covariate selection procedures based on statistical associations only could result in inappropriate adjustment, selection bias or reduction of precision of the exposure effect estimate[35,36].

It has been claimed that post-selection inference cannot be valid at all[6]. Since research on this issue is ongoing[37,38], neutral comparison studies and user-friendly implementations are still lacking[39] and hence its advances are hardly accessible to epidemiologists. Frequentist statistical theory assumes that the parameters to be estimated in a model are fixed before observing the data, while variable selection involves the data in the selection process, meaning the model is not fixed a priori. Consequently, CIs based on the selected model are no longer valid and often underestimate uncertainty in the effect estimator[5,20,37].

Finally, there is no one-size-fits-all implementation of automated covariate selection[5,20] and recommendations on covariate selection may not be applicable to a particular study. Choices regarding covariate selection should strongly depend on the aim of a study, which could be causal inference, prediction, or description[40,41]. Statistical texts that explain variable selection do not always relate implementation of the procedure to those distinct research aims[40].

4 | Simulations

4.1 | Simulation design

Aim: We examined the effect of backward elimination vs. full model specification on the efficiency of causal effect evaluation in simulation studies. First, we performed a proof-of-concept simulation (Experiment 1) to confirm inequality (1). Additionally, we studied the value of backward elimination in efficiency of causal effect estimation in more complex and realistic settings (Experiment 2).

Data-generating mechanisms: The generated data consisted of a binary outcome, Y, a binary exposure, A, and a set of continuous covariates, L. The set of covariates was free of mediators and colliders and was the starting point for all backward elimination procedures. In Experiment 1, the generated data contained a single continuous covariate next to the exposure and outcome. The exposure effect was null, the sample size was set to 60 or 120 and the event fraction (i.e., $Pr(Y = 1)$) was set to 0.5 or 0.2, The conditional

associations *A–L* and *Y–L* varied between 0 and 0.5 on a log-odds scale. A total of 144 scenarios were evaluated. In Experiment 2, the log(cOR) of the exposure was either log(1) or log(1.5). *L* consisted of 24 continuous covariates from a multivariate normal distribution with mean 0 and a variance-covariance matrix with 1s on the diagonal and 0.3 on all off-diagonal elements. The set consisted of a mix of 12 - 24 true confounders, 0 - 12 (near) instrumental variables, 0 - 12 (near) predictors of the outcome, and 0 - 12 noise variables, where the number of each covariate type was varied across simulation scenarios (see Table 2). The expected number of events was set to 50 or 200 and the expected event fraction was set to 0.2 or 0.03, resulting in samples with 250; 1,667; 1,000; or 6,667 observations. Table 2 presents the values of other simulation parameters. A total of 3,960 scenarios were evaluated.

Target estimand: The estimands were the cOR and the mRR of the association between *A* and *Y*, controlled for confounding.

Methods: The cOR was obtained from logistic regression models estimated using Firth's Logistic regression with Intercept Correction (FLIC) to avoid introduction of finite sample bias[26-28] and issues with separation in the simulation[42]. The mRR was estimated using FLIC models that estimated potential outcomes[29-31]. Estimates were evaluated on a logarithmic scale because of the asymmetrical nature of ORs and RRs.

Table 2 Simulation parameters of experiment 2.

Parameter	Value
Conditional exposure-outcome effect	0, log(1.5)
Fixed confounders: conditional log odds ratio confounder-exposure association*	log(1.05)
Fixed confounders: conditional log odds ratio confounder-outcome association*	log(1.05)
Mixture of covariates: conditional log odds ratio covariate-exposure association (4 sets of 3 covariates)*	0, log(1.05), log(1.2)
Mixture of covariates: conditional log odds ratio covariate-outcome association (4 sets of 3 covariates)*	0, log(1.05), log(1.2)
Covariate correlation across all 24 covariates	0.3
Number of events	50, 200
Expected event fraction	0.2, 0.03

* Of the 24 continuous covariates, 12 were assumed to be fixed confounders, and 12 represented a mixture of true confounders (log(1.2)), (near-) instrumental variables (log(1.05)), (near-)predictors of the outcome (log(1.05)) and noise variables (0). In each data set, the number of respective covariate types was determined by the combination of conditional coavariate-exposure/outcome parameters.

Simulations were performed using R statistical software version 3.6.2.[43] using the package `logistf`[44] to implement Firth's correction. In Experiment 1, the MSE of the $\log(\widehat{cOR})$ and $\log(\widehat{mRR})$ was evaluated under two estimation strategies: 'always include covariate L' versus 'always omit covariate L'. In Experiment 2, we evaluated the MSE of the $\log(\widehat{cOR})$ and $\log(\widehat{mRR})$ obtained using a full model versus using backward elimination with cut-off value p = 0.157 (corresponding with using the Akaike information criterion)[24]. We obtained the true mRR for each scenario by a large sample approximation (N = 1,000,000).

Performance measures: The MSE was defined as the average squared difference between the estimated $\log(\widehat{cOR})$ and true log(cOR) or the estimated $\log(\widehat{mRR})$ and true log(mRR) averaged per scenario over the simulation runs (10,000 for Experiment 1; 1,000 for Experiment 2). We compared the full and selected model in terms of relative efficiency of the $\log(\widehat{cOR})$ and $\log(\widehat{mRR})$, which was computed as a ratio of the MSE obtained from the backward elimination procedure divided by the MSE obtained from the full model.

This simulation design was reported following previous recommendations[45]. All R code for simulations is available at https://github.com/Kluijken/CI_CovSel.

4.2 | Simulation results

Experiment 1: Inequality (1) held for most (90%) of the simulated scenarios in Experiment 1 for the $\log(\widehat{cOR})$ and for 20% of the scenarios regarding the $\log(\widehat{mRR})$. Hence, regarding the cOR, omitting the covariate was often more beneficial in terms of MSE than including it (see Online Supplementary File 2). Regarding the mRR, including the covariate was often more beneficial in terms of MSE than omitting it (Figure 2). Omitting the covariate was beneficial in terms of MSE only when the covariate was an instrument or a near-instrument. The benefit of omitting was larger when the event fraction was lower, 0.2 instead of 0.5, and, as expected, when sample size was lower, 60 compared to 120.

Experiment 2: In Experiment 2, the median relative efficiency of the $\log(\widehat{cOR})$ across all scenarios was 1.04, indicating the MSE was on average lower for the full model compared to the selected model. The median relative efficiency of the $\log(\widehat{mRR})$ across all scenarios was 1.05. Across all 3,960 scenarios, the bias of the full model was zero for both the $\log(\widehat{cOR})$ and$\log(\widehat{mRR})$, whereas the average bias across all scenarios of the backward

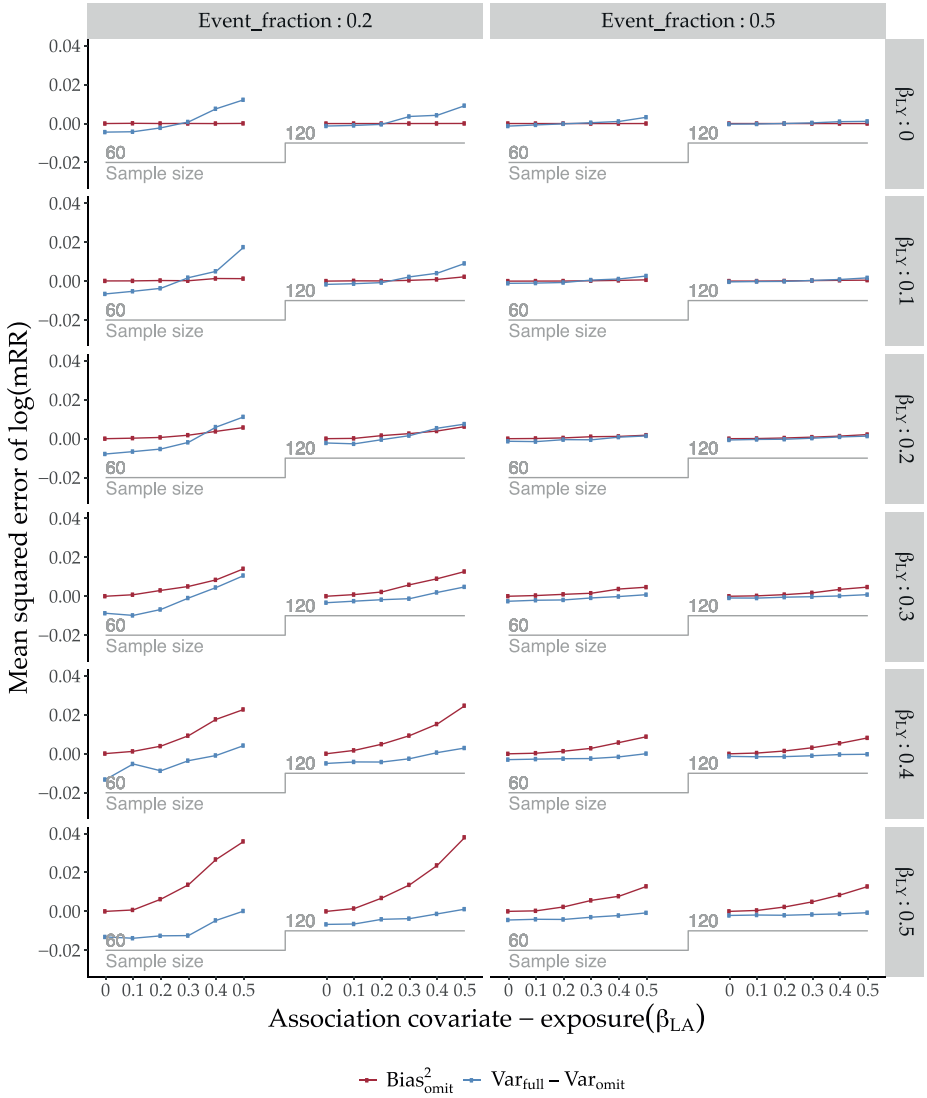


Figure 2. Results of simulation Experiment 1 for the marginal risk ratio (mRR). A single covariate L acts as a confounder, (near-)instrumental variable, (near-)predictor of the outcome, or noise variable in a setting where a binary exposure has a true null effect on a binary outcome. The squared bias and difference in variance is compared when L is always included or always omitted, illustrating principle (1) in the main text. β_{LA} and β_{LY} refer to the conditional log odds ratio of the covariate-exposure and covariate-outcome association, respectively. This figure was created using the `looplot` package[46].

eliminated model was 0.03 for the $\log(\widehat{cOR})$ and 0.02 for the $\log(\widehat{mRR})$, compared to 0.56 and 0.52 for $\log(\widehat{cOR})$ and $\log(\widehat{mRR})$, respectively, for a model without any covariate adjustment. We found 112 scenarios (2.8%) for cOR and 47 scenarios (1.1%) for mRR in which the MSE was lower for the selected than the full models.

Closer examination of the 112 scenarios in which the $(\widehat{cOR})$ estimated using backward elimination showed lower MSE than the full model revealed that 100 scenarios included at least three full instrumental variables and 37 scenarios included at least three noise variables (see Table 3). In these scenarios, the increased efficiency remained small, with a minimal relative efficiency of 0.97, meaning that the MSE when backward elimination was applied was only 3% lower than the MSE of the full model in the most beneficial setting. For the $\log(\widehat{mRR})$, we found that 42 scenarios included at least three full instrumental variables and 18 scenarios included at least three noise variables (see Table 4). Again, the increased efficiency remained small, with a minimal relative efficiency of 0.97. Full results of the simulations are presented in Online Supplementary File 3.

5 | Discussion

Our simulation results show that, compared to estimating a model with all prespecified confounders, application of backward elimination was unlikely to reduce the MSE of the exposure effect estimator (defined as the cOR and the mRR), while introducing a bias. We identified some settings in which the MSE of the effect estimators was lower with backward elimination than without, yet the reduction in MSE was small. The results are driven by two antagonist effects; an MSE-reducing effect of omitting weak confounders, and an MSE-increasing effect caused by additional uncertainties incurred by applying automated selection as explained by Heinze et al (2018).

Our findings support and extend previous recommendations on automated covariate selection. VanderWeele and Shpitser proposed to use the disjunctive cause criterion for confounder selection[15]. This criterion states to control for covariates that are either a cause of the exposure or a cause of the outcome, which may lead to adjustment for instrumental variables. Therefore, they recommended to implement backward elimination or forward selection to eliminate such variables. Our findings provide weak support for the use of variable selection in this case. In an overview and classification of covariate selection strategies, Witte and Didelez found that backward elimination performed well in terms of bias in the effect estimator in settings that

Table 3 Summary of simulation experiment 2; results for the conditional odds ratio (cOR). Each row represents 495 scenarios with varying associations between the covariates and the exposure and/or outcome. Mean bias indicates the average bias of the log(cOR) for the full and backward eliminated model, respectively. Relative efficiency of the mean squared error (MSE) of the cOR is computed as a ratio of the backward elimination MSE divided by the full model MSE.

Conditional exposure effect	Event fraction	Number of events	Mean bias full	Mean bias BE	Median relative efficiency	Minimum relative efficiency	Maximum relative efficiency	Number of scenarios MSE(cOR) BE < full	At least 3 IVs in DGM	At least 3 noise variables in DGM	No IVs or noise in DGM
0	0.20	50	0.00	0.04	1.08	1.00	1.20	0	0	0	0
0	0.20	200	0.00	0.02	1.03	0.98	1.13	10	9	3	0
0	0.03	50	0.00	0.03	1.03	0.98	1.10	6	6	1	0
0	0.03	200	0.00	0.02	1.02	0.97	1.08	42	40	14	0
log(1.5)	0.20	50	0.00	0.06	1.09	1.02	1.27	0	0	0	0
log(1.5)	0.20	200	0.00	0.03	1.04	0.98	1.13	8	8	3	0
log(1.5)	0.03	50	0.00	0.04	1.04	0.99	1.10	7	5	3	1
log(1.5)	0.03	200	0.00	0.02	1.02	0.98	1.09	39	32	13	3
Overall results			0.00	0.03	1.04	0.97	1.27	112	100	37	4

Abbreviations: BE = backward elimination, cOR = conditional odds ratio, DGM = data-generating mechanism, IV = instrumental variable, MSE = mean squared error.

Table 4 Summary of simulation Experiment 2; results for the marginal risk ratio (mRR). Each row represents 495 scenarios with varying associations between the covariates and the exposure and/or outcome. Mean bias indicates the average bias of the log(mRR) for the full and backward eliminated model, respectively. Relative efficiency of the mean squared error (MSE) of the mRR is computed as a ratio of the backward elimination MSE divided by the full model MSE.

Conditional exposure effect	Event fraction	Number of events	Mean bias full	Mean bias BE	Median relative efficiency	Minimum relative efficiency	Maximum relative efficiency	Number of scenarios MSE(mRR) BE < full	At least 3 IVs in DGM	At least 3 noise variables in DGM	No IVs or noise in DGM
0	0.20	50	0.00	0.03	1.12	1.06	1.25	0	0	0	0
0	0.20	200	0.00	0.02	1.05	1.00	1.17	0	0	0	0
0	0.03	50	0.00	0.03	1.05	1.00	1.12	0	0	0	0
0	0.03	200	0.00	0.02	1.02	0.97	1.09	21	19	8	0
log(1.5)	0.20	50	-0.02	0.03	1.12	1.05	1.31	0	0	0	0
log(1.5)	0.20	200	0.00	0.02	1.05	1.00	1.15	1	1	0	0
log(1.5)	0.03	50	-0.01	0.03	1.05	1.01	1.12	0	0	0	0
log(1.5)	0.03	200	0.00	0.02	1.02	0.98	1.10	25	22	10	0
Overall results			0.00	0.02	1.05	0.97	1.31	47	42	18	0

Abbreviations: BE = backward elimination, DGM = data-generating mechanism, IV = instrumental variable, mRR = marginal risk ratio, MSE = mean squared error.

contained strong confounders and instrumental variables and did not perform well when applied to a sufficient adjustment set in which each confounder was responsible for a small degree of confounding[17]. We found similar patterns in terms of the MSE of the effect estimator, irrespective of conditional or marginal effects are of interest. On the other hand, Vansteelandt and colleagues recommended against the use of automated covariate selection even when there is a potential efficiency gain by excluding an instrumental variable, because this would prevent overstating the precision with which a causal effect is known[18]. Summarizing, the true number of irrelevant covariates and instruments included in the prespecified set of adjustment variables, and the strength of association of true confounders with the outcome greatly affect the relative performance of applying backward elimination. In practice, these conditions are usually unknown, but the more domain expertise is available to define the set, the less a researcher has to rely on data-driven selection.

Our motivating example was typical for clinical observational studies where a set of covariates is available that accurately describes the health state of a subject just before the decision to perform an intervention or not, but where dependencies among these covariates are difficult to assess. Therefore, we only assumed that the set of covariates was free of mediators and that there was no unmeasured confounding. These assumptions were based on clinical expertise and allowed specification of an initial working set without explicitly specifying a full DAG. Under these conditions, backward elimination was applied to potentially increase the efficiency of the effect estimate by setting weak covariate effects to zero, but not to change the underlying assumptions.

A limitation of our study is that we did not consider scenarios in which clinical expertise is not available. In many clinical settings, it is questionable whether the assumption of no residual confounding really holds. Furthermore, it is difficult to judge to what extent preselection can be reliably done. This depends on the novelty of a research field and often one will rely on previous research to derive assumptions. Doing so, researchers should be aware of inappropriate methodology, such as questionable conclusions stemming from observed bivariate associations, which typically do not reflect multivariable relations represented in a causal network[36]. It is up to the researcher to explain to what extent pre-processing based on background knowledge is possible and hence whether data-driven selection could be of added value.

Additionally, as our paper was intended to evaluate a common practice, we did not consider more sophisticated approaches for data-driven confounder selection. Whereas backward elimination is an outcome-oriented selection procedure, other approaches,

such as Lasso penalized regression approaches[47,48], take into account both covariate-outcome and covariate-exposure relations. Such approaches might lead to more robust and efficient effect estimation compared to backward elimination; however, they are hardly ever used in epidemiological studies. We also excluded augmented backward elimination[16] and other novel approaches as we were either involved in developing these methods or lack the necessary expertise to apply them routinely.

We conclude that backward elimination for confounder selection is unlikely to have added value when an initial set of covariates related to the exposure and/or outcome can be specified based on background knowledge. If researchers choose to perform backward elimination of potential confounders, selection should be justified, e.g., because a large number of potential confounders are anticipated to function as (near-)instruments, and the approach should be prespecified in a statistical analysis plan. Covariate selection based solely on statistical criteria should be avoided due to the possible selection of mediators and colliders. Irrespective of whether or not covariate selection strategies are being applied, we recommend to always provide information about the assumed causal structure, ideally by a depiction of assumed causal dependencies, but at least by excluding mediators and the possibility of unmeasured confounding.

Appendix

Mean squared error of ordinary least squares exposure effect estimator of full and reduced model

Notation and set-up

Consider the model depicted in Figure A1. Let A denote the exposure, L a covariate and Y the outcome. Each variable is a linear combination of the variables affecting it (indicated by the directed arrows in Figure 1) plus an error term. The coefficients of the model are denoted α for the relation between A and L, γ for the relation between Y and L conditional on A and β for the relation between Y and A conditional on L. All variables are normally distributed, where $L \sim \mathcal{N}(\mu_L, \sigma_L^2)$, $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$, and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.

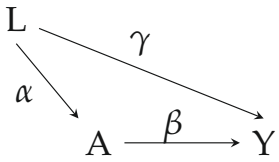


Figure A1: Directed acyclic graph.

To explore when elimination of a covariate improves efficiency of an ordinary least squares estimator of the association between exposure and the outcome (in short exposure effect estimator), we derive expressions for the MSE of the regression coefficient of Y on A when L is included in the model (full model), $\text{MSE}(\hat{\beta})$, and for the model where L is omitted (reduced model), $\text{MSE}(\hat{\beta}_{omit})$. The expression for bias of the exposure effect estimator is derived using path-tracing rules, as described by Wright[49] and Pearl[50]. The variance of the exposure effect estimator is specified for an ordinary least squares estimator in a finite sample with n observations, similar to the expressions for the extended omitted variable framework by Cinelli and Hazlett[51].

Full model

Consider an ordinary least squares regression of the model in Figure A1,

$$Y = \hat{\beta}A + \hat{\gamma}L + \hat{\epsilon},$$

where Y is an n x 1 vector containing the outcome of interest for each of the n observations, A is an $n \times 1$ vector of the continuous exposure variable, L is an $n \times 1$ vector of the continuous covariate, $\hat{\beta}$ and $\hat{\gamma}$ are ordinary least squares coefficients of the association between Y and A and between Y and L, respectively, and $\epsilon \sim \mathcal{N}(0, \sigma^2_{Y^{\perp A,L}})$. Let $Y^{\perp A,L}$ denote the variable Y after removing the components linearly explained by A and L, $A^{\perp L}$denote the variable A after removing the components linearly explained by L, $\hat{\alpha}$ denote the ordinary least squares estimator of the association between A and L and $\text{var}(\cdot)$ and $\text{cov}(\cdot)$ denote the sample variances and covariances, respectively[51]. Then, using path-tracing rules[49,50]

$$\begin{aligned}
\text{var}(L) &:= \hat{\sigma}^2_L \\
\text{var}(A) &:= \hat{\sigma}^2_A \\
\text{cov}(A, L) &:= \hat{\sigma}^2_L \hat{\alpha} \\
\text{var}(A^{\perp L}) &:= \hat{\sigma}^2_A - \hat{\sigma}^2_L \hat{\alpha}^2 \\
\text{var}(Y) &:= \hat{\sigma}^2_Y \\
\text{cov}(Y^{\perp L}, A^{\perp L}) &:= (\hat{\sigma}^2_A - \hat{\sigma}^2_L \hat{\alpha}^2)\hat{\beta} \\
\text{var}(Y^{\perp A,L}) = \hat{\sigma}^2_{Y^{\perp A,L}} &:= \hat{\sigma}^2_Y - \hat{\sigma}^2_A \hat{\beta}^2 - \hat{\sigma}^2_L \hat{\gamma}^2 - 2\hat{\sigma}^2_L \hat{\alpha}\hat{\beta}\hat{\gamma}.
\end{aligned}$$

As described in[49,50], partial regression coefficients can be readily read from a path diagram such as Figure A1. The expected value for the partial regression coefficient regressing Y on A given L, $\hat{\beta}$, can be expressed as

$$\begin{aligned}
\hat{\beta} &= \frac{\text{cov}(Y^{\perp L}, A^{\perp L})}{\text{var}(A^{\perp L})}, \text{ and} \\
\mathbb{E}(\hat{\beta}) &= \frac{(\sigma^2_A - \sigma^2_L \alpha^2)\beta}{\sigma^2_A - \sigma^2_L \alpha^2} \\
&= \beta.
\end{aligned} \tag{1}$$

Hence, the bias of the effect estimator in the full model is $\mathbb{E}(\hat{\beta}) - \beta = 0$.

Let df denote the ordinary least squares regression's degrees of freedom. An expression for the variance of $\hat{\beta}$ is[51]

$$\begin{aligned}\text{var}(\hat{\beta}) &= \frac{\text{var}(Y^{\perp A,L})}{\text{var}(A^{\perp L})}\frac{1}{\text{df}-1} \\ &= \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_A^2\hat{\beta}^2 - \hat{\sigma}_L^2\hat{\gamma}^2 - 2\hat{\sigma}_L^2\hat{\alpha}\hat{\beta}\hat{\gamma}}{\hat{\sigma}_A^2 - \hat{\sigma}_L^2\hat{\alpha}^2}\frac{1}{n-3}.\end{aligned}$$

To derive the expected variance of $\hat{\beta}$, we use the property that $\hat{\sigma}^2 \approx \hat{\sigma}^2\frac{n-1}{n} \approx \sigma^2$ for $\hat{\sigma}_A^2$, $\hat{\sigma}_L^2$ and $\hat{\sigma}_Y^2$. Furthermore, we assume that the residuals $\sigma_{Y^{\perp A,L}}^2$ are independent from A and L and apply a first-order approximation[52]. Then,

$$\mathbb{E}[\text{var}(\hat{\beta})] \approx \frac{\sigma_Y^2 - \sigma_A^2\beta^2 - \sigma_L^2\gamma^2 - 2\sigma_L^2\alpha\beta\gamma}{\sigma_A^2 - \sigma_L^2\alpha^2}\frac{1}{n-3}. \tag{2}$$

Since $MSE = bias^2 + variance$,

$$MSE(\hat{\beta}) \approx \frac{\sigma_Y^2 - \sigma_A^2\beta^2 - \sigma_L^2\gamma^2 - 2\sigma_L^2\alpha\beta\gamma}{\sigma_A^2 - \sigma_L^2\alpha^2}\frac{1}{n-3}. \tag{3}$$

.

Reduced model

Consider the ordinary least squares regression of the model in Figure 1 where variable L is omitted

$$Y = \hat{\beta}_{omit}A + \hat{\epsilon}_{omit},$$

where Y is an $n \times 1$ vector containing the outcome of interest for each of the n observations, A is an $n \times 1$ continuous exposure variable, $\hat{\beta}_{omit}$ is an ordinary least squares coefficient estimator and $\epsilon_{omit} \sim \mathcal{N}(0, \sigma_{Y^{\perp A}}^2)$. Let $Y^{\perp A}$ denote the variable Y after removing the components linearly explained by A. Then,

$$\begin{aligned}\text{cov}(Y, A) &:= \hat{\sigma}_A^2\hat{\beta} + \hat{\sigma}_L^2\hat{\alpha}\hat{\gamma} \\ \text{var}(Y^{\perp A}) = \hat{\sigma}_{Y^{\perp A}}^2 &:= \hat{\sigma}_Y^2 - \hat{\sigma}_A^2\hat{\beta}^2 - 2\hat{\sigma}_L^2\hat{\alpha}\hat{\beta}\hat{\gamma}.\end{aligned}$$

Again, reading partial regressions from Figure A1 as described in [49,50], we find that the expected value for the marginal association between A and Y can be expressed as

$$\begin{aligned}\hat{\beta}_{omit} &= \frac{\text{cov}(Y, A)}{\text{var}(A)}, \text{ and} \\ \mathbb{E}(\hat{\beta}_{omit}) &= \frac{\sigma_A^2\beta + \sigma_L^2\alpha\gamma}{\sigma_A^2} \\ &= \beta + \frac{\sigma_L^2\alpha\gamma}{\sigma_A^2}.\end{aligned} \tag{4}$$

Hence, the expected bias in the reduced model is $\mathbb{E}(\hat{\beta}_{omit}) - \beta = \frac{\sigma_L^2 \alpha\gamma}{\sigma_A^2}$. We obtain the expression for the variance of $\hat{\beta}_{omit}$[51]

$$\begin{aligned}\text{var}(\hat{\beta}_{omit}) &= \frac{\text{var}(Y^{\perp A})}{\text{var}(A)} \frac{1}{\text{df}} \\ &= \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_A^2 \hat{\beta}^2 - 2\hat{\sigma}_L^2 \hat{\alpha}\hat{\beta}\hat{\gamma}}{\hat{\sigma}_A^2} \frac{1}{n-2}.\end{aligned}$$

To derive the expected variance of $\hat{\beta}_{omit}$, we use the property that $\hat{\sigma}^2 \approx \hat{\sigma}^2 \frac{n-1}{n} \approx \sigma^2$ for $\hat{\sigma}_A^2$ and $\hat{\sigma}_Y^2$. Furthermore, we assume that the residuals $\sigma_{Y^{\perp A}}^2$ are independent from A and apply a first-order approximation[52]. Then,

$$\mathbb{E}[\text{var}(\hat{\beta}_{omit})] \approx \frac{\sigma_Y^2 - \sigma_A^2 \beta^2 - 2\sigma_L^2 \alpha\beta\gamma}{\sigma_A^2} \frac{1}{n-2}, \tag{5}$$

where the final step is performed using a first-order approximation[52] and under the assumption that the residuals $\sigma_{Y^{\perp A}}^2$ are independent from A. Since $MSE = bias^2 + variance$,

$$\begin{aligned}MSE(\hat{\beta}_{omit}) &\approx \left(\frac{\sigma_L^2 \alpha\gamma}{\sigma_A^2}\right)^2 + \frac{\sigma_Y^2 - \sigma_A^2 \beta^2 - 2\sigma_L^2 \alpha\beta\gamma}{\sigma_A^2} \frac{1}{n-2} \\ &\approx \frac{\sigma_L^4 \alpha^2 \gamma^2}{\sigma_A^4} + \frac{\sigma_Y^2 - \sigma_A^2 \beta^2 - 2\sigma_L^2 \alpha\beta\gamma}{\sigma_A^2} \frac{1}{n-2}.\end{aligned} \tag{6}$$

Comparison full and reduced model

Plotting equation (3) and equation (6) for the arbitrarily chosen values $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.2$, $\sigma_A^2 = 2.5$, $\sigma_L^2 = 8$, and $\sigma_Y^2 = 10$ yields the following result.

The lines in Figure A2 show that $\text{MSE}(\hat{\beta}_{omit}) < \text{MSE}(\hat{\beta})$ for smaller sample size n, and that the full model has a lower MSE for $\hat{\beta}$ with larger sample sizes. In other words, if the inequality $Bias_{omit}^2 < Variance_{full} - Variance_{omit}$ holds, i.e., when

$$\frac{\sigma_L^4 \alpha^2 \gamma^2}{\sigma_A^4} < \frac{\sigma_Y^2 - \sigma_A^2 \beta^2 - \sigma_L^2 \gamma^2 - 2\sigma_L^2 \alpha\beta\gamma}{\sigma_A^2 - \sigma_L^2 \alpha^2} \frac{1}{n-3} - \frac{\sigma_Y^2 - \sigma_A^2 \beta^2 - 2\sigma_L^2 \alpha\beta\gamma}{\sigma_A^2} \frac{1}{n-2}$$

holds, then the reduced variance by omitting L outweighs the increase in squared bias. A simple approximation of the value of n for which this is the case, denoted n', can be

found by assuming $n-3 \approx n-2 \approx n$ (which is reasonable to assume for sufficiently large sample sizes). This yields the following expression

$$n' < \frac{\sigma_A^4\sigma_Y^2 - \sigma_A^6\beta^2 - \sigma_A^4\sigma_L^2\gamma^2 - 2\sigma_A^4\sigma_L^2\alpha\beta\gamma}{(\sigma_A^2 - \sigma_L^2\alpha^2)\sigma_L^4\alpha^2\gamma^2} - \frac{\sigma_A^2\sigma_Y^2 - \sigma_A^4\beta^2 - 2\sigma_A^2\sigma_L^2\alpha\beta\gamma}{\sigma_L^4\alpha^2\gamma^2}. \quad (7)$$

Similar to the plot above, Equation (7) indicates that for sample sizes smaller than the critical n', omitting L results in a lower mean squared error of the exposure estimate than including it. The impact of many of the parameters in the equation depends on the values of other parameters. Their joint effect on the critical sample size could be assessed by means of plotting the result of equation for various parameters values, as is presented in Figures A3 and A4.

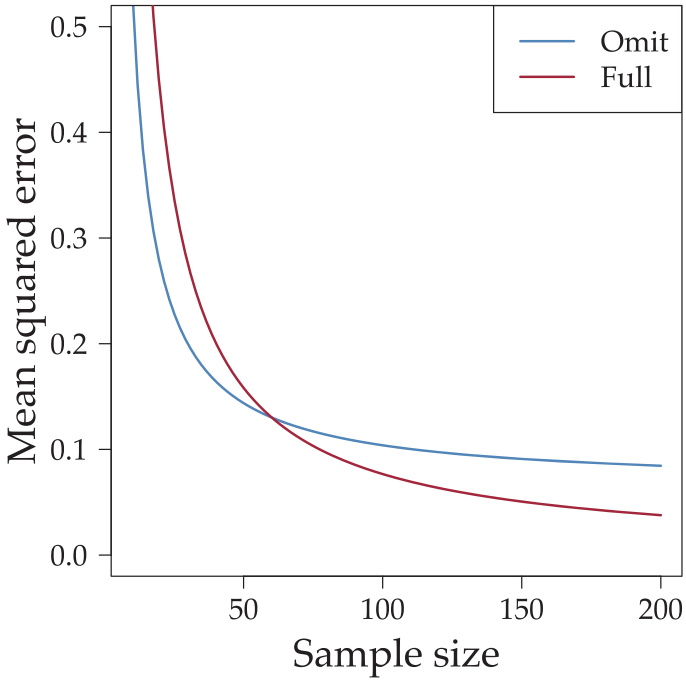


Figure A2. Illustration of the bias-variance trade-off for the ordinary least squares estimator of the exposure effect when including (Full) or omitting (Omit) covariate *L*. The red and blue line are computed using expression (3) and (6), respectively, for sample size n ranging from 0 to 200 and the values $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.2$, $\sigma_A^2 = 2.5$, $\sigma_L^2 = 8$, and $\sigma_Y^2 = 10$. The value of n for which the reduced variance by omitting L outweighs the increase in squared bias is around 60.

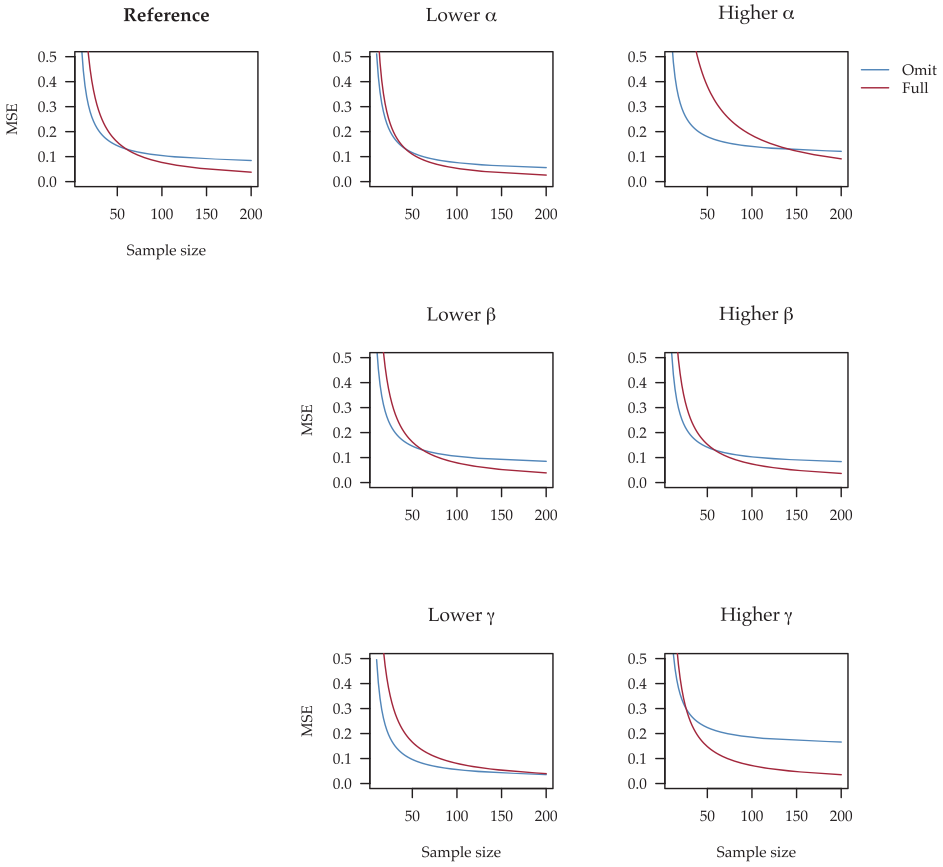


Figure A3. Illustration of the bias-variance trade-off for the ordinary least squares estimator of the exposure effect when including (Full) or omitting (Omit) covariate *L*. The red and blue line are computed using expression (3) and (6), respectively, for sample size n ranging from 0 to 200. The parameters σ^2_A, σ^2_L, and σ^2_Y are fixed to values 2.5, 8 and 10, respectively, while the values of parameters α, β, and γ are varied. The reference plot is created using the arbitrarily chosen values $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.2$, $\sigma^2_A = 2.5$, $\sigma^2_L = 8$, and $\sigma^2_Y = 10$. Lower α indicates $\alpha = 0.3$ and higher $\alpha = 0.5$. Lower β indicates $\beta = 0.2$ and higher $\beta = 0.4$. Lower γ indicates $\gamma = 0.1$ and higher $\gamma = 0.3$.

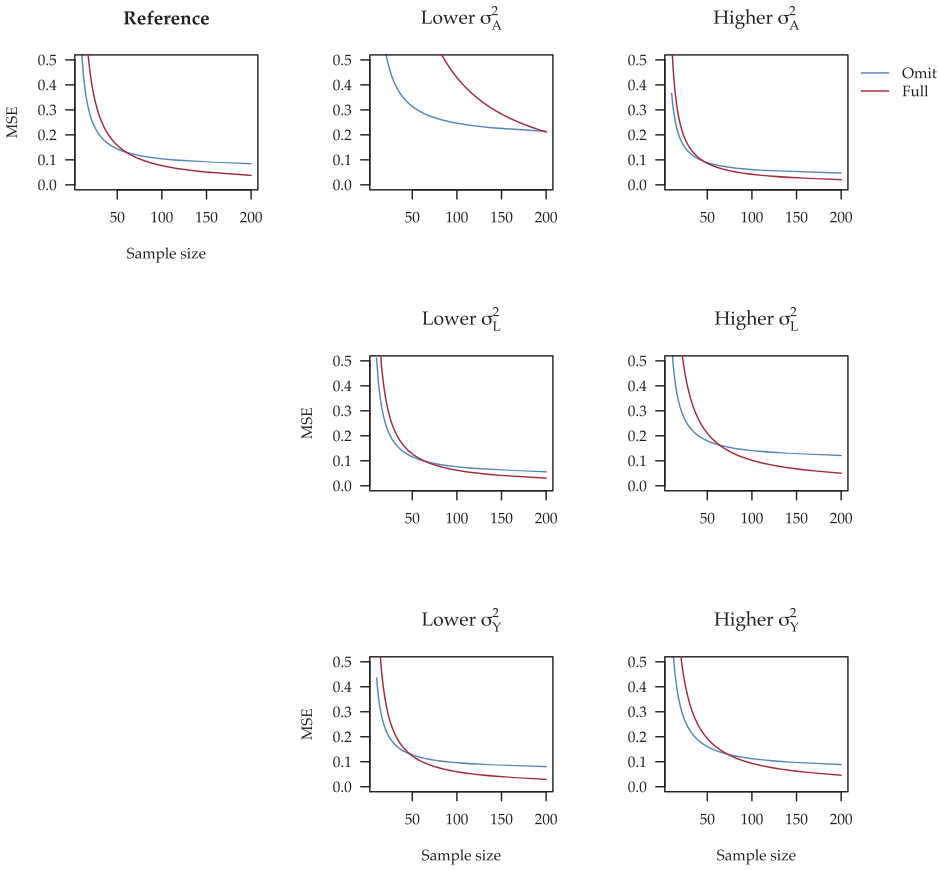


Figure A4. Illustration of the bias-variance trade-off for the ordinary least squares estimator of the exposure effect when including (Full) or omitting (Omit) covariate *L*. The red and blue line are computed using expression (3) and (6), respectively, for sample size n ranging from 0 to 200. The parameters α, β, and γ are fixed to values 0.4, 0.3 and 0.2, respectively, while the values of parameters σ^2_A, σ^2_L , and σ^2_Y are varied. The reference plot is created using the arbitrarily chosen values $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.2$, $\sigma^2_A = 2.5$, $\sigma^2_L = 8$, and $\sigma^2_Y = 10$. Lower σ^2_A indicates $\sigma^2_A = 1.5$ and higher $\sigma^2_A = 3.5$. Lower σ^2_L indicates $\sigma^2_L = 6$ and higher $\sigma^2_L = 10$. Lower σ^2_Y indicates $\sigma^2_Y = 8$ and higher $\sigma^2_Y = 12$.

Online Supplementary Files

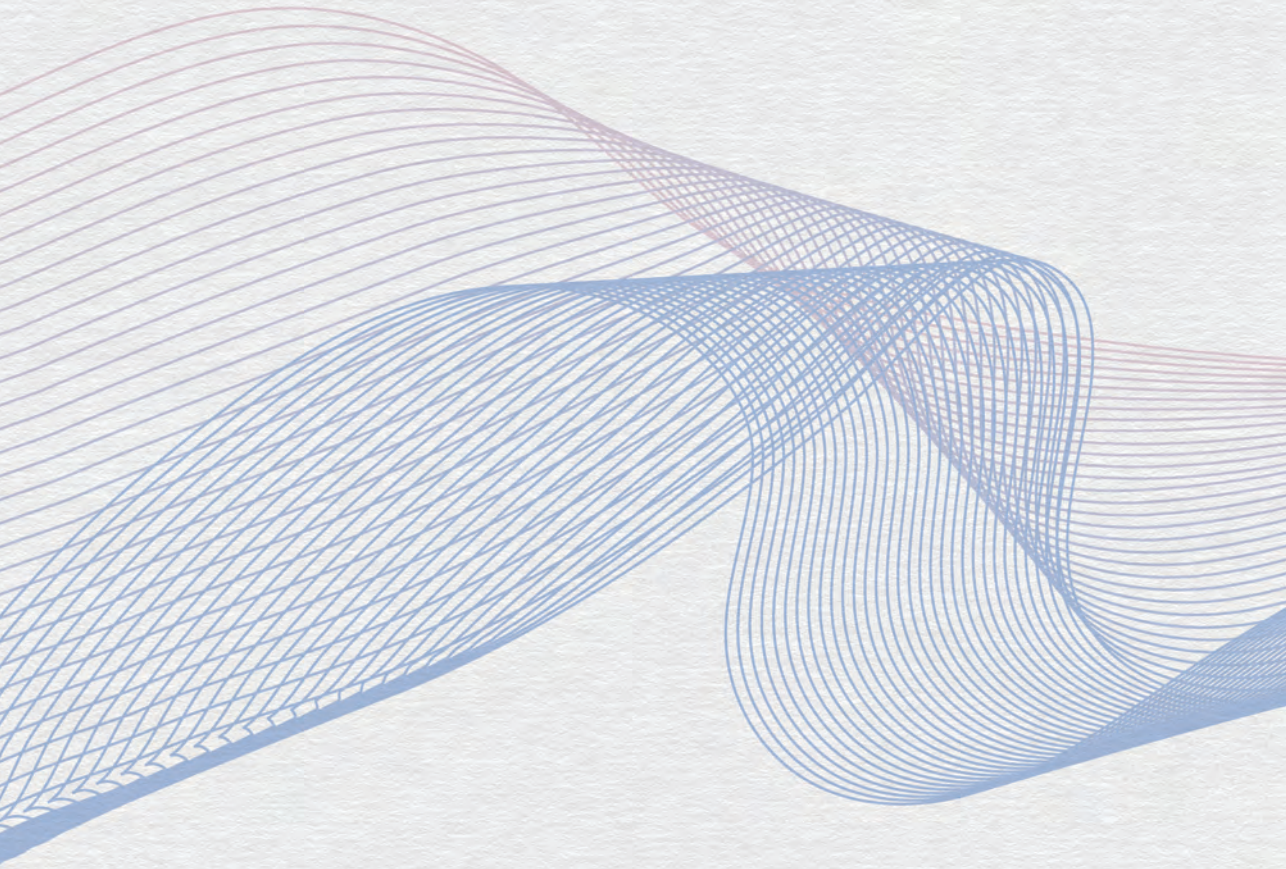
The supplementary files referred to in this Chapter are available online at https://github.com/KLuijken/Dissertation_Online_Supplements/tree/main/Chapter_4

References

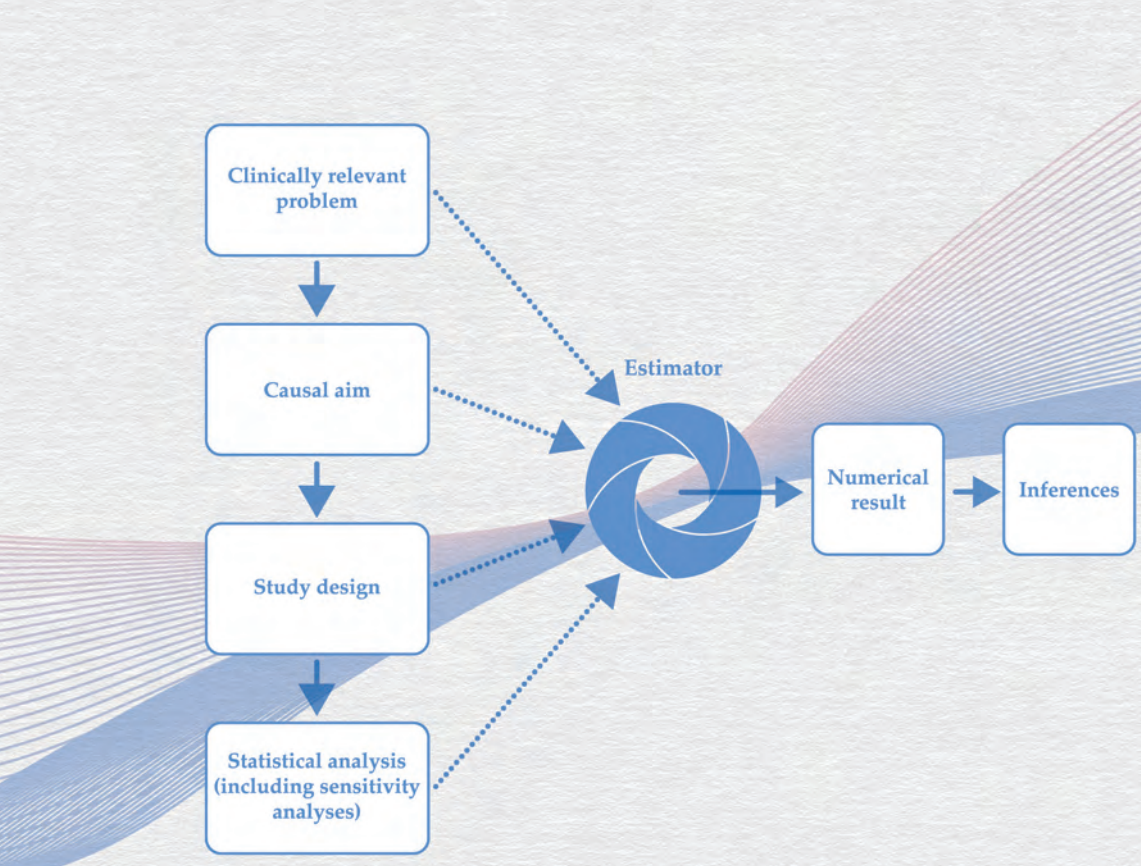
1. Hernán MA, Robins JM. *Causal inference: what if*. In: Boca Raton: Chapman & Hall/CRC; 2020.
2. VanderWeele TJ. Principles of confounder selection. *European Journal of Epidemiology*. 2019;34(3):211-219.
3. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003;14(3):300-306.
4. Ding P, Miratrix LW. To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. *Journal of Causal Inference*. 2015;3(1):41-57.
5. Heinze G, Wallisch C, Dunkler D. Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*. 2018;60(3):431-449.
6. Leeb H, Pötscher BM. Model selection and inference: Facts and fiction. *Econometric Theory*. 2005:21-59.
7. Talbot D, Massamba VK. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. *European Journal of Epidemiology*. 2019;34(8):725-730.
8. Groenwold RH, Van Deursen AM, Hoes AW, Hak E. Poor quality of reporting confounding bias in observational intervention studies: a systematic review. *Annals of Epidemiology*. 2008;18(10):746-751.
9. Ali MS, Groenwold RH, Belitser SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of Clinical Epidemiology*. 2015;68(2):122-131.
10. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. *European Journal of Epidemiology*. 2009;24(12):733-736.
11. Pouwels KB, Widyakusuma NN, Groenwold RH, Hak E. Quality of reporting of confounding remained suboptimal after the STROBE guideline. *Journal of Clinical Epidemiology*. 2016;69:217-224.
12. Klein-Geltink J, Rochon P, Dyer S, Laxer M, Anderson G. Readers should systematically assess methods used to identify, measure and analyze confounding in observational cohort studies. *Journal of Clinical Epidemiology*. 2007;60(8):766. e1-11.
13. Hemkens LG, Ewald H, Naudet F, et al. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. *Journal of Clinical Epidemiology*. 2018;93:94-102.
14. Greenland S, Daniel R, Pearce N. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *International Journal of Epidemiology*. 2016;45(2):565-575.
15. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics*. 2011;67(4):1406-1413.
16. Dunkler D, Plischke M, Leffondré K, Heinze G. Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. *PloS one*. 2014;9(11):e113677.
17. Witte J, Didelez V. Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal*. 2019;61(5):1270-1289.
18. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*. 2012;21(1):7-30.
19. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology*. 2008;167(5):523-529.
20. Sauerbrei W, Perperoglou A, Schmid M, et al. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. *Diagnostic and Prognostic Research*. 2020;4:1-18.
21. Sandner SE, Nolz R, Loewe C, et al. Routine preoperative aortic computed tomography angiography is associated with reduced risk of stroke in coronary artery bypass grafting: a propensity-matched analysis. *European Journal of Cardio-Thoracic Surgery*. 2020;57(4):684-690.
22. Gregorich MG. *A comparison of methods for causal inference with a rare binary outcome* [Master thesis (unpublished)], Wien; 2018.
23. Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I, Topic group Causal Inference (TG7) of the STRATOS initiative. Formulating causal questions and principled statistical answers. *Statistics in medicine*. 2020;39(30):4922-4948.

24. Harrell Jr FE. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer; 2015.
25. Tennant P, Murray E, Arnold K, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology*. 2020;dyaa213.
26. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in medicine*. 2002;21(16):2409-2419.
27. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38.
28. Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Statistics in Medicine*. 2017;36(14):2302-2317.
29. Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *Journal of Clinical Epidemiology*. 2010;63(1):2-6.
30. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American Journal of Epidemiology*. 2004;160(4):301-305.
31. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of Clinical Epidemiology*. 2007;60(9):874-882.
32. Wallisch C, Dunkler D, Rauch G, De Bin R, Heinze G. Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling. *Statistics in Medicine*. 2021;40(2):369-381.
33. Schaefer RL. Bias correction in maximum likelihood logistic regression. *Statistics in Medicine*. 1983;2(1):71-78.
34. Cordeiro GM, McCullagh P. Bias correction in generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1991;53(3):629-643.
35. Heinze G, Dunkler D. Five myths about variable selection. *Transplant International*. 2017;30(1):6-10.
36. Sun G-W, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology*. 1996;49(8):907-916.
37. Berk R, Brown L, Buja A, Zhang K, Zhao L. Valid post-selection inference. *The Annals of Statistics*. 2013;41(2):802-837.
38. Belloni A, Chernozhukov V, Wei Y. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*. 2016;34(4):606-619.
39. Kammer M, Dunkler D, Michiels S, Heinze G. Evaluating methods for Lasso selective inference in biomedical research by a comparative simulation study. In. *arXiv preprint arXiv:2005.07484*2020.
40. Shmueli G. To explain or to predict? *Statistical Science*. 2010;25(3):289-310.
41. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *Chance*. 2019;32(1):42-49.
42. van Smeden M, de Groot JA, Moons KG, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*. 2016;16(1):1-12.
43. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020.
44. Heinze G, Ploner M, Jiricka L. *logistf: Firth's Bias-Reduced Logistic Regression*. R package version 1.24. 2020 https://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-software/fllogistf/
45. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*. 2019;38(11):2074-2102.
46. Kammer M. *looplot: Create nested loop plots*. R package version 0.5.0.9001. 2020 https://github.com/matherealize/looplot/
47. Ertefaie A, Asgharian M, Stephens DA. Variable selection in causal inference using a simultaneous penalization method. *Journal of Causal Inference*. 2018;6(1).

48. Wilson A, Reich BJ. Confounder selection via penalized credible regions. *Biometrics*. 2014;70(4):852-861.
49. Wright S. Correlation and causation. *Journal of Agricultural Research*. 1921;20:557-85.
50. Pearl J. Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference*. 2013;1(1):155-170.
51. Cinelli C, Hazlett C. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2020;82(1):39-67.
52. Henckel L, Perković E, Maathuis MH. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:190702435*. 2019.



5


Clinically relevant problem
Causal aim
Study design
Statistical analysis (including sensitivity analyses)
Estimator
Numerical result
Inferences

How to assess applicability and methodological quality of studies of operative interventions in orthopedic trauma surgery

It is challenging to generate and subsequently implement high quality evidence in surgical practice. A first step towards improving the quality of surgical studies is to assess current methods to grade the strengths and weaknesses of surgical evidence and appraise current risk of bias tools for the surgical community. Here, we described items that are common to different risk-of-bias tools, how these could be used to assess operative intervention studies in orthopedic trauma surgery, and how these relate to applicability of results. We extracted information from the Cochrane risk-of-bias-2 (RoB-2) tool, Risk Of Bias In Non-randomised Studies - of Interventions tool (ROBINS-I), and methodological index for non-randomized studies (MINORS) criteria and derived a concisely formulated set of items tailored to operative interventions in orthopedic trauma surgery. The set contained nine items: population, intervention, comparator, outcome, confounding, missing data and selection bias, intervention status, outcome assessment, and pre-specification of analysis. Each item can be assessed using signaling questions and was explained using good practice examples of operative intervention studies in orthopedic trauma surgery. The set of items will be useful to form a first judgment on studies that have been included in a systematic review. Existing risk of bias tools can be used for further evaluation of methodological quality. Additionally, the proposed set of items might be a helpful starting point for peer reviewers.

1| Background

It is challenging to generate and subsequently implement high quality evidence in surgical practice[1]. In the field of orthopedic trauma surgery, it takes approximately ten years from design to execution of an RCT[2]. What is more, Oberkofler and colleagues showed that results of surgical RCTs often do not convince the surgical community of their findings due to a substantial risk of bias[3]. This is a highly undesirable situation, because a lot of effort, time, public money, and patient participation is spent on research with futile impact on surgical care[4].

To what extent a study can inform surgeons and patients depends on its applicability and methodological quality. Appraising the methodological quality of a study and judging the applicability (external validity or generalizability) of study results to clinical practice remains challenging in the field of surgical research, especially the assessment of bias (internal validity). This is reinforced by the fact that systematic reviews of operative interventions increasingly include both randomized controlled trials (RCTs) and observational studies[5,6], adding to the complexity of the assessment.

Many comprehensive risk-of-bias tools are available to assess the methodological quality of studies of interventions[7-10], such as the Cochrane risk-of-bias (RoB 2) tool for randomized trials[11] and the Risk Of Bias In Non-randomised Studies - of Interventions (ROBINS-I) tool[12]. However, these tools focus on internal validity (risk of bias) aspects of a study and do not simultaneously evaluate clinical applicability of the results. The tools were often developed with a focus on studies of pharmacological interventions and may therefore not be ideally suited for studies of operative interventions. Additionally, it is convenient to assess both types of studies using a single list of items.

Here, we describe the selection of items that are common to different risk-of-bias tools, how these could be used to assess operative intervention studies in orthopedic trauma surgery, and how these relate to applicability of results. We take the perspective of a researcher who performs a systematic review and wants to make a first judgment on the applicability and methodological quality of included studies. Relevance of these items for editors, peer reviewers, and researchers will be addressed in the discussion section.

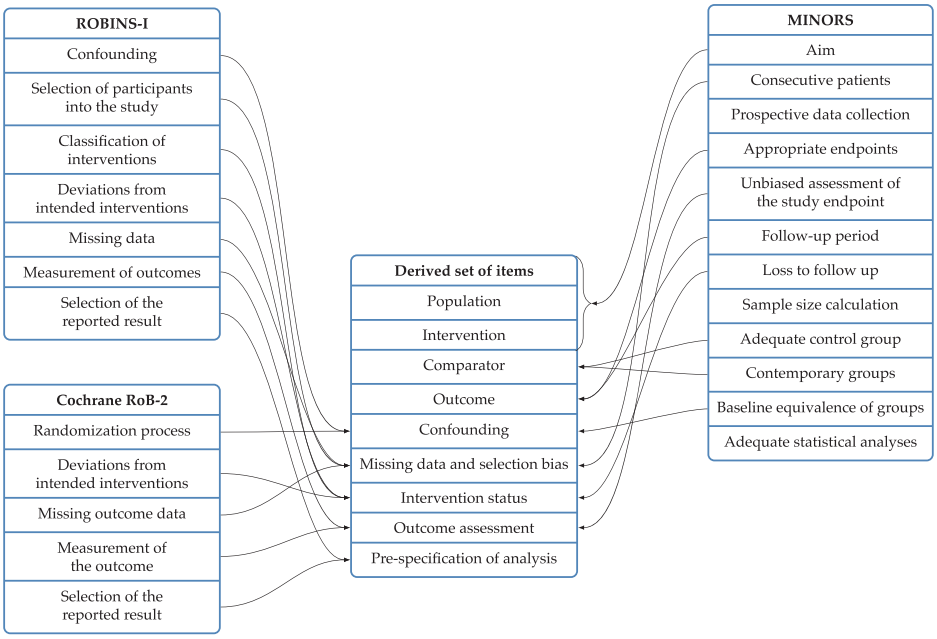


Figure 1. Flow diagram indicating which existing risk of bias tool the signaling questions in the concise set were based on.

2| Extracting items from existing risk-of-bias tools

To establish an easy-to-use set of items for assessment of applicability and methodological quality of studies of operative interventions, we extracted information from the RoB-2[11], ROBINS-I[12], and methodological index for non-randomized studies (MINORS) criteria[13] and derived a concisely formulated set of items tailored to operative interventions. It has been pointed out that the methodologically rigorous RoB-2 and ROBINS-I tools require a high level of statistical knowledge, making their implementation challenging and time consuming[14-16]. We aimed to summarize scoring items in such a way that the items were easy to use for assessment of articles of both RCTs and observational studies of operative interventions.

All signaling questions from the RoB-2, ROBINS-I, and MINORS were taken as a starting point. We identified signaling questions with overlapping topics. Based on their relevance to surgical studies, the overlapping set formed the initial key items. We then evaluated remaining signaling questions. Questions that were less relevant for studies of operative interventions, such as questions regarding time-varying exposures,

Table 1 Overview of items and overarching questions of the established set. The Appendix describes the signaling questions for each item.

Applicability		
Item	**Overarching question**	**Explanation**
Population	Is the patient population included in the study representative of the patient population defined in the PICO of the systematic review?	Patients included in the study are ideally representative of patients that would be typically encountered in the clinical practice setting for which the PICO is defined.
Intervention	Is the investigated intervention representative of the intervention defined in the PICO of the systematic review?	The studied operative intervention is ideally performed similar to the procedure that would be typically performed in the clinical practice setting for which the PICO is defined.
Comparator	Is the comparator intervention representative of the comparator defined in the PICO of the systematic review?	The comparator intervention is ideally performed similar to the procedure that would be typically performed in the clinical practice setting for which the PICO is defined.
Outcome	Is the outcome representative of the outcome defined in the PICO of the systematic review?	The outcome should be relevant to patients typically encountered in the clinical practice setting for which the PICO is defined and should be measured using an appropriate procedure at the appropriate time.
Methodology		
Item	**Overarching question**	**Explanation**
Confounding	Is there comparability of intervention groups, or are appropriate methods applied to correct for incomparability?	An important assumption needed to make causal claims about effects of operative interventions is comparability of intervention groups. This can be established through appropriate randomization (in randomized studies) or adjustment for confounding (in observational studies).
Missing data and selection bias	Were the patients included in the analysis representative of all patients included in the study and was the impact of missing data negligible?	Selection of patients based on missing values or phenomena that occurred after inclusion into the study can introduce bias in the effect of the operative intervention.
Intervention status	Was the intervention status correctly classified?	To infer the effect of a operative intervention it should be clear how crossovers and deviations from planned interventions are dealt with in the analysis.
Outcome assessment	Was the outcome correctly measured?	The study outcome should be measured such that it does not influence the estimated effect by (dis)favoring the outcome of one of the intervention groups.
Pre-specification of analysis	Were analyses prespecified and did the study adhere to the specified analysis plan?	Specifying analyses prior to analyzing the data prevents researchers from (often unintentionally) trying many approaches to fit the data, defining the research question/ hypothesis only after observing the result, and selectively reporting the findings that yield the desired result.

were discarded and questions were reformulated to be more appropriate for a surgical context (Figure 1). The set of items was further improved by user experiences in an accompanying study that assessed the applicability and methodological quality of studies from two recent systematic reviews (van de Wall, forthcoming).

The finally established set contained four items on applicability and five items of methodology (Table 1). Each item contained multiple signaling questions that help to arrive at an overarching judgement of the study quality regarding that topic, with some signaling questions specifically applicable only to RCT or observational studies. The proposed set aimed to summarize key information needed to assess the applicability and methodological quality of operative intervention studies, but the choice for items and signaling questions was surely arbitrary, and the set is opened to further elaborations and improvements.

3| Nine items to assess applicability and methodological quality

The set contained nine items: population, intervention, comparator, outcome, confounding, missing data and selection bias, intervention status, outcome assessment, and pre-specification of analysis. This set can be split in two subsets, representing applicability (first four items) and methodological quality (remaining five items).

3.1 | Items of applicability

The first four items represent the starting point of almost every clinical study, which is a clearly articulated research question (see Box 1). In a systematic review, the research question determines which original studies should be included as well as the degree to which they can provide valuable evidence. The well-known PICO acronym can be a helpful structure when defining a research question about the possible effect of an operative intervention[17].

Box 1. Well-definedness of research questions is crucial in studies of complex interventions.

Studies investigating causal effects of interventions, both randomized and non-randomized, provide scientific evidence to inform medical decisions about those interventions. Ideally, a study indicates clearly which medical decision can be informed by the findings by unambiguously defining the research question, specifying the target population, the intervention strategies that are compared, and what outcome is considered (and when).

In studies of pharmacological interventions, a research question could for instance be 'what is the effect of taking drug A compared to taking drug B on a particular outcome in a specific population?' Although this seems trivial, some parts of this question are not yet clearly defined. How is the drug administered (e.g., oral, or intravenous) and what dosages are compared? Other aspects, however, may be irrelevant, such as the hand with which a pill was taken or what shoes the individual was wearing when they took the drug. A research question should be sufficiently well defined in the sense that all *relevant* aspects are specified and thus should be addressed in the study design and analysis[18,19].

Arguably, pharmacological interventions consist of less components than operative interventions and it is more straightforward to define them precisely. Studies of operative interventions go beyond a mere description of surgical techniques; other relevant aspects include the pre- and post-surgery treatment, experience of the surgeon and team, and more. On top of that, the operative intervention itself is tailored to a particular patient[20]. Hence, defining all relevant aspects in a research question demands considerable time and effort in studies of operative interventions.

For further reading on sufficiently well-defined research questions, we refer to[18] and[19].

As an example, consider the PICO for the systematic review on operative treatment of proximal humerus fractures in the accompanying paper by van de Wall and colleagues (van de Wall, forthcoming). The PICO of the systematic review was to compare functional outcomes measured using a validated functional score for the shoulder one year after plate osteosynthesis (minimally invasive or open reduction and internal fixation) followed by 6 weeks none-weightbearing functional treatment versus one year

after initiation of conservative intervention, consisting of 6 weeks of no weight bearing, pain-guided movement and a sling if necessary, in patients with a closed, displaced, proximal humerus fractures older than 18 years.

Item 1. Population

The population defined in a research question ideally matches the patient population typically encountered in the clinical setting for which the study is conducted. In orthopedic trauma surgery, elements that define the population are, e.g., the anatomical location of the fracture, the type of fracture (e.g., open/closed, simple/multifragmentary, or combination), and age group. Fjalestad and colleagues[21] defined the relevant population as "*patients aged 60+ years with a displaced, unstable three-or four-part proximal humerus fracture of OTA group 11-B2 or 11-C2 (displaced fracture of extra-articular or articular, bifocal type) without previous shoulder injuries*". Because the population of interest was clearly reported (and its characteristics summarized in a table), the degree to which it matches to the population specified in the example PICO of the systematic review can easily be assessed.

Item 2. Intervention

Obviously, the studied operative intervention should be clearly defined. In case of an operative intervention, this entails, e.g., specification of the osteosynthesis material, surgical approach and the type and duration of the post-operative treatment regime. In case of a conservative intervention, the duration and type of conservative intervention should be clearly reported.

For example, Fjalestad and colleagues[21] defined the studied intervention as follows: "*Patients allocated to surgery were operated on within 1 week of hospital admission. The goal of surgery was anatomic reduction of the fracture and fracture stabilization [using angular stable plate] to allow for early mobilization. After surgery, patients were immobilized in a modified Velpeau bandage until self-exercises and training instructed by a physical therapist were started on the third postoperative day.*" This was accompanied by a detailed account of the operative technique and the physiotherapy protocol, such that it was clear from the description what the intervention constituted. The intervention corresponds to the intervention defined in the example PICO of the systematic review, with the exception that the post-treatment regime was extended to include strengthening exercises after 6 weeks and a recommendation of physical therapy for at least 6 months.

Other relevant aspects of the intervention are whether study hospitals routinely perform the intervention, which help to clarify whether participating surgeons are experienced in conducting the investigated procedure. For instance, Fjalestad and colleagues indicated that: "*Three surgeons performed all operations and were trained in the surgical technique before performing surgery on study participants. Surgeons 1, 2, and 3 performed 18, five, and two operations, respectively. Surgery occurred during daytime hours*"[21]. Also, a learning curve (or the absence thereof) could be relevant. For example, Knobe et al. compared helical blade nailing of the femoral head versus locked plating and reported that[22]: "*[t]hree surgeons […] were proficient in the locked plating technique and three […] were proficient with helical blade nailing. Both implants had been used by the surgeons for more than 3 years, so they would have been beyond the learning curve and they had a comparable experience level for each implant*".

Item 3. Comparator

Similar to the studied intervention, the comparator intervention should be clearly defined, and the same considerations apply. For example, Fjalestad and colleagues[21] defined the comparator intervention as follows: "*On admission to the hospital, patients were immobilized in a modified Velpeau bandage. All patients allocated to conservative treatment stayed in the hospital for at least 1 day and received the same instructions from the physiotherapist as patients allocated to surgery*", accompanied by a description of an optional closed reduction procedure. The unambiguous reporting of the conservative treatment regime allowed for assessment of the applicability of the comparator arm with respect to the comparator specified in the example PICO of the systematic review. While the conservative intervention is roughly similar to the definition of the comparator intervention in the systematic-review PICO, the optional closed reduction was not part of the systematic-review PICO.

Item 4. Outcome

Specification of a relevant study outcome consists of three parts: the outcome definition, the timepoint at which the outcome is assessed and the measurement procedure or instrument by which the outcome is assessed. For example, Fjalestad and colleagues[21] defined the primary outcome as functional outcome at one year, indicating the outcome definition and timepoint at which it was assessed. The outcome measurement was the Constant score, which is a score ranging 0 – 100 measured by self-reported pain (max. 15 points), self-reported activities of daily-living (max. 20 points), range of

motion (forward and lateral elevation, max. 10 points each, and external and internal rotation, max. 10 points each), and power (25 points)[23]. The unambiguous reporting of the outcome definition, timepoint and measurement procedure allowed for assessment of the applicability of the outcome with respect to the outcome specified in the example PICO of the systematic review.

3.2 | Items of methodology

Five methodological items are key for assessing methodological quality of a study: confounding, missing data and selection bias, classification of intervention status, outcome assessment, and pre-specification of the statistical analysis. Each of the items will be discussed below.

Item 5. Confounding

Comparability of intervention groups is essential for evaluation of effects of operative interventions and can be invoked by appropriate randomization (in randomized studies) or adjustment for confounding (in observational studies).

In randomized studies, a random allocation sequence and concealment of that allocation contribute to comparability of intervention groups, leading to comparability in observed (and unobserved) characteristics of study groups at baseline. An example of a clear description of the randomization procedure is given by Rangan and colleagues[24]: "*After obtaining informed consent and key baseline information, research associates randomly allocated patients to surgical or nonsurgical treatment using an independent remote randomization service (telephone or online access) provided by the York Trials Unit (University of York). Randomization was performed using a computer program with 1:1 allocation, stratifying by tuberosity involvement (yes or no) and using random block sizes of 4, 8, and 12.*" Based on this information, it can be assessed that the allocation sequence was random. Furthermore, inspection of baseline differences between intervention groups suggested no clinically relevant differences in observed characteristics. "*The baseline characteristics […] for randomized patients (N = 250) and those providing [Oxford Shoulder Score] data at 2 years (n = 215) were well balanced except for smoking status (there were more smokers in the nonsurgical group)*"[24].

Ideally, the allocation sequence is concealed at least until patients are enrolled in the study[25]. In case research associates or patients are aware which intervention the next enrolled patient will receive this might influence the decision to enroll (the patient) into

the study and thus limit comparability of study groups. Hence, a detailed description of the allocation procedure is needed to assess the validity of the intervention allocation.

In observational studies, allocation of intervention is no random process, and intervention groups cannot be presumed to be comparable. Therefore, a key requirement for observational studies of operative interventions is that researchers argue convincingly that intervention groups are comparable or that they provide enough detail to assess whether important clinical characteristics are sufficiently controlled for in the statistical analysis of the study[26].

An example of the former is a study by Beks and colleagues, who compared the effect of rib fixation based on a clinical treatment algorithm on intensive care unit length of stay to nonoperative intervention for both patients with a flail chest and patients with multiple rib fractures[27]. They compared groups of patients with rib fractures admitted to hospitals that either operated most patients or mostly treated patients conservatively. Allocation of emergency patients to hospitals is to a certain extent a random process, based on availability and location of the accident. When different hospitals treat patients with similar symptoms with different interventions, this allows for a natural experiment by comparing outcomes across hospitals[28]. In this example, confounding due to severe incomparability of intervention groups was deemed unlikely by design. Additionally, Beks and colleagues adjusted for a number of confounders using propensity score matching.

Indeed, when intervention groups cannot be considered to be (fully) comparable by design, statistical adjustment for measured confounders can be considered. For example, Jenkinson and colleagues adjusted for variables that are considered to be confounders, because they are known risk factors of the outcome and/or they may have contributed to the indication for a particular intervention[29]: "*The factors considered to be the most important confounders also contributing to deep-infection risk were chosen for the propensity-score algorithm. These factors included patient age, sex, time delay to debridement, fracture grade (Gustilo-Anderson grade I, II, or IIIA), evidence of gross contamination, tibial compared with nontibial site, and ASA class (1 or 2 compared with 3 or higher). These factors were chosen, based on consensus among the investigators, as the factors most important for predicting later infection but also as those most divergent between the immediate and delayed-closure groups*". Jenkinson and colleagues selected confounders based on background knowledge, in line with recommendations that specialist knowledge about the relation

between covariates and the complex intervention and/or outcome is needed to identify a set of potential confounders.

A common misconception is that confounders can be identified based on statistical criteria. In fact, statistical criteria cannot identify nor discard covariates as being confounder variables[30-34]. Of note, most statistical methods to adjust for confounding (including propensity score methods) can only adjust for measured confounding variables. After confounding adjustment, bias due to unmeasured confounding may still be present, e.g. because a confounder was measured inaccurately (or a continuous variable was dichotomized), or not measured at all[35]. A final note on confounders is that it is advisable not to interpret coefficients of confounding variables as causal effects or independent prognostic associations[36].

Item 6. Missing data and selection bias

Data are often incomplete. In some circumstances, data can be missing without substantially affecting the results. When this is the case, a study report should clarify why missingness is thought to have no effect on the study outcome, as was done for example by Portinary and colleagues[37]: "*To evaluate the impact of the emergency operations on postoperative functional status, the [activities of the daily living (ADL)] scores at the time of discharge were compared to the pre-admission ADL scores using the Chi-square test. Only patients for whom both pre-admission and postoperative ADL scores were available were included in this analysis. The subgroup analysis comparing patients with missing ADL score data with those where data was available showed no differences in terms of demographic and baseline characteristics […]. Therefore, participants without missing ADL score data were considered as a random sample of the study population. Therefore, missing data were considered to be completely at random and a complete case analysis was performed*".

In many cases, however, excluding patients for whom information on some variables is missing can introduce bias, because the missingness is related to observed or unobserved characteristics of the patients[38-40]. Beks and colleagues assumed missingness in their study was at random and described how it was dealt with[27]: "*We applied multiple imputation (25 times) to impute missing values for ASA [2.1% (7/332)], TTSS [20% (67/332)], AIS head [0.6% (2/332)], pulmonary contusion [0.6% (2/332)], pH [9.0% (30/332)], and base excess [9.0% (30/332)]. Multiple imputation was performed using the mice() algorithm in R*".

Studies should describe patterns of missing data and describe the assumed missing data mechanism. Otherwise, it is impossible to assess the potential impact of missing data and whether this was dealt with appropriately. The validity of performing a

complete case analysis cannot be assessed from a study that merely states that patients with missing values were excluded from analysis. Pointing out that few cases were missing is not a valid justification of complete case analysis, since the proportion of missing data is not directly linked to the severity of the bias that is introduced by it[41].

Apart from variables having missing values, subjects can also be missing entirely in case they are not included in the study, which could lead to selection bias. However, if those included in the study are representative of the entire set of eligible subjects, the risk of this type of bias seems small. Klei and colleagues provided a clear description why patients included in the analyses seemed representative of all patients included in the study[42]: "*Among the 116 sternovertebral fracture patients, 43 patients were excluded from further analysis (1 military patient, 14 patients who died early after admission before fracture treatment, 14 patients with either isolated upper cervical spine or lower lumbar spine fractures, and 14 patients who were lost to follow-up). The remaining 73 patients were included for further analysis*".

Sometimes patients are excluded from analysis because they do not consent to participate in the study. A comparison between patients that consented and refused to partake in the study can be done to assess the possibility that selection bias is introduced, as is described by Rangan and colleagues[24]: "*Of the 563 eligible patients, 250 (44%) consented to take part in the trial […]. The mean age of the […] participants was 66 years (range, 24-92 years), 192 (77%) were female, and 249 (99.6%) were white. These characteristics were similar to patients who refused consent (mean age, 68 years; 75% female).*"

Item 7. Intervention status

The defined PICO specifies which operative interventions are compared including their post-intervention regimens. However, deviations from these ideally unambiguously defined, yet potentially hypothetical, situations can occur in clinical practice, both in the intervention and comparator arm.

The intervention status can be incorrectly registered in the data, referred to as 'misclassification'. For instance, when data are retrieved from electronic health records, the procedure may be inaccurately registered or incorrectly extracted into the analytical dataset. A patient may falsely be recorded not to have received an operative intervention, while they actually had, or selection bias can be introduced in case of a comparison of operative interventions. However, in most cases, misclassification of surgical interventions seems unlikely.

Defining the intervention status of a patient in the final analysis is not straightforward when the patient was assigned to one intervention arm, but actually received the opposite intervention (too). This is commonly referred to as a *cross-over*. Which intervention status patients should then be assigned to depends on the aim of the study, and in particular on the intervention effect of interest. For instance, an RCT by Van der Meijden and colleagues aimed to estimate an intention-to-treat effect and assessed patients in the intervention group that they were randomized to[43]. "*One patient (2%) in the plate group and six patients (10%) in the nailing group underwent intraoperative crossover to the other treatment group and were further analyzed as part of their original treatment group according to the intention-to-treat principle*". Consequently, the result of the study no longer represents an effect of plate fixation versus intramedullary nailing on functional recovery. Rather, it represents an effect of plate fixation with optional revision using intramedullary nailing versus intramedullary nailing with optional revision using plate fixation on functional recovery. Although this interpretation is arguably less straightforward, it might be the effect of main interest in clinical practice.

Considerations regarding patients' intervention status differ slightly between RCTs and observational studies. In RCTs, a cross-over commonly refers to a patient who was assigned to a particular intervention, but then received an alternative intervention, meaning that the patient received a single intervention. In observational studies, a cross-over commonly refers to a patient who received a particular intervention first and then received the alternative intervention, meaning that the patient received both interventions. Cross-over interventions in comparisons of operative versus non-operative interventions often pose a more challenging problem than crossovers between operative interventions.

Finally, including the post-operative treatment regime for determining a patient's intervention status likely complicates matters considerably. Adherence to post-operative treatment is often less well documented and post-operative treatment options may be combined for some patients.

Item 8. Outcome assessment

Ideally, the study outcome is measured in the same way in all study patients, notably irrespective of the intervention a patient received. This can be achieved by means of a valid and reliable procedure to measure the outcome[44]. For instance, the outcome 'quality of life' can be measured using a well-established questionnaire such as the EQ5D, as was done by Banierink and colleagues[45]: "*Quality of life was assessed with the*

EuroQol 5D (EQ-5D). The EQ-5D is a brief questionnaire that measures health-related quality of life based on five dimensions of health: mobility, self-care, usual activities, pain/discomfort and anxiety/depression [17]."

Functional outcomes are ideally measured using validated instruments, too, as was done by Ochen and colleagues [46]: "*Functional outcome was assessed at least 12 months following [the operative intervention], using the Dutch language version of the QuickDASH score. The QuickDASH is a validated and shortened version of the Disabilities of the Arm, Shoulder and Hand questionnaire (DASH)*".

When the outcome is measured using a non-standardized measurement, outcome values may be less reliable. For instance, when forward or lateral elevation of the shoulder is measured by visual inspection rather than by use of a goniometer, values may be less reliable and inter-rater variability is likely increased. On top of that, an outcome assessor may (subconsciously) be affected by knowledge about the intervention that a patient received (i.e., when they are unblinded). These considerations apply to functional outcomes and self-reported outcomes, including patient reported outcome measures (PROMs), alike.

To prevent bias by unblinded outcome assessment, Nauth and colleagues designed their RCT anticipating the bias that could be introduced by differential unblinded outcome assessment of their primary outcome re-operation[47]: "*Surgeons and patients were not blinded. However, we did minimise the associated risk of bias with central and independent, although unblinded, radiographic adjudication of the primary endpoint*". A Committee adjudicated re-operations at the end of follow-up, where re-operation was defined as surgery to promote fracture healing, relieve pain, treat infection, or improve function within 24 months after the initial procedure (described in detail in the supplements of [47]).

Item 9. Pre-specification of analysis

The credibility of results can be diminished by trying many approaches to fit the data and selectively reporting the results that yield the desired outcome. When the choice to perform a statistical test depends on patterns in the data, the expected number of false positives (i.e., type I error rate) is likely inflated[48]. Similarly, the type I error rate increases when more statistical tests are conducted on the same data set, thus performing multiple statistical tests without reporting all of them in the published manuscript prohibits readers from assessing the potential for false positive findings. Although such data dredging and cherry picking has harmful consequences, these practices may well be conducted unintentionally – especially when findings (in

hindsight) are convincing and easy to explain. To overcome this problem, statistical data analysis should be prespecified as much as possible, e.g., by means of a statistical analysis plan that defines which analyses will be performed and the methods used to perform these analyses, including handling of missing data[49]. For RCTs, preregistration of the study protocol is considered the norm[50], but for observational studies, study protocols seem to be pre-specified less often, although the urgency to do this is certainly recognized[51].

To further enhance transparency, protocols can be made publicly available to allow for assessment of protocol adherence. Protocols can be preregistered at, e.g., https://clinicaltrials.gov/ (for RCTs), https://www.isrctn.com/ (both RCTs and observational studies), https://osf.io/ (both RCTs and observational studies), and protocols for systematic reviews can be preregistered on https://www.crd.york.ac.uk/prospero/ or https://osf.io. Journals such as International Journal of Surgery Protocols or the British Medical Journal Open allow for publication of study protocols.

Good examples of publicly available study protocols are a trial by Smeeing and colleagues, who compared functional outcome twelve weeks after randomization to unprotected non-weight-bearing, protected weight-bearing, or unprotected weight-bearing as tolerated in patients who underwent surgical fixation of ankle fractures[52]. The protocol is available at https://www.trialregister.nl/, NTR3727[53]. Taha and colleagues registered an observational pilot study to assess the feasibility of performing an RCT to study the effect of operative intervention of metacarpal fractures affecting the index to little finger(s) compared to non-operative intervention. The study is currently ongoing and is registered at ISRCTN (13922779).

Red flags

Apart from the aforementioned items, a study can contain aspects that set alarm bells ringing about the quality of the methodology or statistical analysis that do not fit within items 5 – 9 specifically. As a systematic reviewer or peer reviewer, it is important to be aware of this and make notes about such red flags.

4| Discussion

We proposed a concise set of items, based on existing risk-of-bias tools, to perform an initial assessment of the applicability and methodological quality of randomized and non-randomized studies into effects of operative interventions in orthopedic trauma surgery. In terms of the IDEAL Framework[54-58], this set of items is intended to assess stage 3 (assessment) and stage 4 (long-term monitoring) studies. This assessment can be done as part of a systematic review to discard studies of low quality with relative ease and to separate out higher quality studies for further scrutiny of methodological quality using available assessment tools[11,12,59,60].

In the current study, we took the perspective of a systematic reviewer, who can use the set of items to appraise studies included in a systematic review and to determine which articles can be considered for a subsequent meta-analysis. However, the proposed set of items might be a helpful starting point also when taking on different roles (Figure 2). The set of items can serve as a reference when peer reviewing an article or when informing medical decisions or policy. While the set of items is primarily derived for assessment of study reports (e.g., manuscript or published articles), it could be perceived as a starting point for researchers when they set up a study or when they report on their own research. However, given the many considerations involved in study conduct, it is advisable to consult other resources when working out a study design and analysis plan.

As systematic reviews of operative procedures increasingly include both RCTs and observational studies[5,6], it is convenient to evaluate both study types with the same set of items. Although RCTs have been described as being more internally valid than observational studies, as reflected in the traditional pyramid of evidence, it becomes increasingly apparent that randomization by design alone is insufficient as a surrogate for risk of bias[6,28,61], and revisions of the pyramid of evidence have been proposed[62]. Including both RCTs and observational studies in systematic reviews is advisable since they potentially provide complementary evidence on the effect of the studied operative intervention.

We intended to establish a set of assessment items that is easy to use with minimal loss of accuracy of the evaluation. The RoB-2 and ROBINS-I have been criticized for being time-consuming and requiring in-depth statistical knowledge, which would hinder their implementation in systematic reviews[14-16]. However, there is an evident trade-off in ease of use and rigor of the assessment. Uptake of rigorous assessment tools can be

improved both by raising awareness and training in the use of available material[63] as well as by making the existing material more accessible. Our proposal is a first step towards bridging intelligibility and scrupulosity in assessment of studies of operative interventions. We encourage further development of an assessment instrument tailored to studies of operative interventions, in particular by bringing together surgical and methodological expertise. In light of such further developments, we point out that studies of operative interventions face a methodological challenge because most studies evaluate complex interventions, but the current set of items does not explicitly address how to evaluate issues introduced by evaluations of complex interventions.

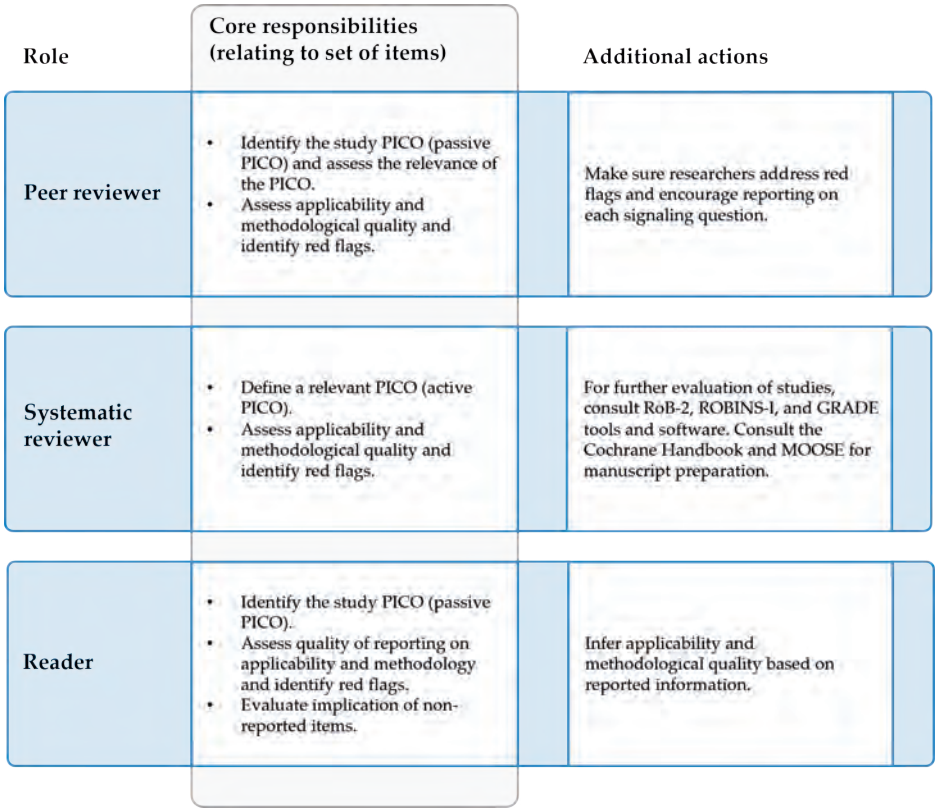


Figure 2. Schematic summary of how the concise set of items can be used by peer reviewers, systematic reviewers, and other readers appraising studies of operative interventions. The contributed value of the set of items ranges from helpful instrument to mere starting point depending on the role of the assessor. When reporting on a study it can be useful to take into account that the study can be read from these perspectives.

In an accompanying study, the set of items was applied to re-assess studies that were included in two published systematic reviews of interventions for proximal humerus factures, providing an illustration of how the proposed items can be used for assessment applicability and methodological quality of randomized and non-randomized studies into effects of operative interventions (van de Wall, forthcoming).

To conclude, the concise set of items can be used for an initial assessment of the applicability and methodological quality of randomized and non-randomized studies into effects of operative interventions. We make a call to use this set not only when performing a systematic review and meta-analysis, but to use it as a reference also when peer reviewing an article, informing medical decisions or policy, or reporting on original research.

Appendix

This file describes items and signaling questions for an assessment of applicability and methodological quality of studies of operative interventions in orthopedic trauma surgery. Researchers can decide on the scoring options for each signaling question, such as yes/no/no information or yes/possibly yes/no/possibly no/no information. Where possible, we recommend documenting quotes that explicitly address a signaling question.

PICO of the systematic review

Population: _____

Intervention: _____

Comparator: _____

Outcome: _____

Applicability

Item	Question
Population	1. Is the patient population included in the study representative of the patient population defined in the PICO of the systematic review?
	1.1. Did inclusion criteria match the patient population specified in the PICO?
	1.2. Was a relevant subgroup of participants excluded?
Intervention	2. Is the investigated intervention representative of the intervention defined in the PICO of the systematic review?
	2.1. Was the investigated intervention similar to the intervention as defined in the PICO?
	2.2. Were the participating surgeons experienced in conducting the investigated procedure?
	2.3. Was the post-operative treatment regime in the intervention arm similar to the one defined in the PICO?
Comparator	3. Is the comparator intervention representative of the comparator defined in the PICO of the systematic review?
	3.1. Was the comparator similar to the comparator as defined in the PICO?
	3.2. Were the health care professionals experienced in conducting the comparator procedure?
	3.3. Was the post-intervention treatment regime in the comparator arm similar to the one defined in the PICO?
Outcome	4. Is the outcome representative of the outcome defined in the PICO of the systematic review?
	4.1. Was the outcome measurement similar to the outcome as defined in the PICO?
	4.2. Was the timing of the outcome described and similar to the specification in the PICO?

Methodology

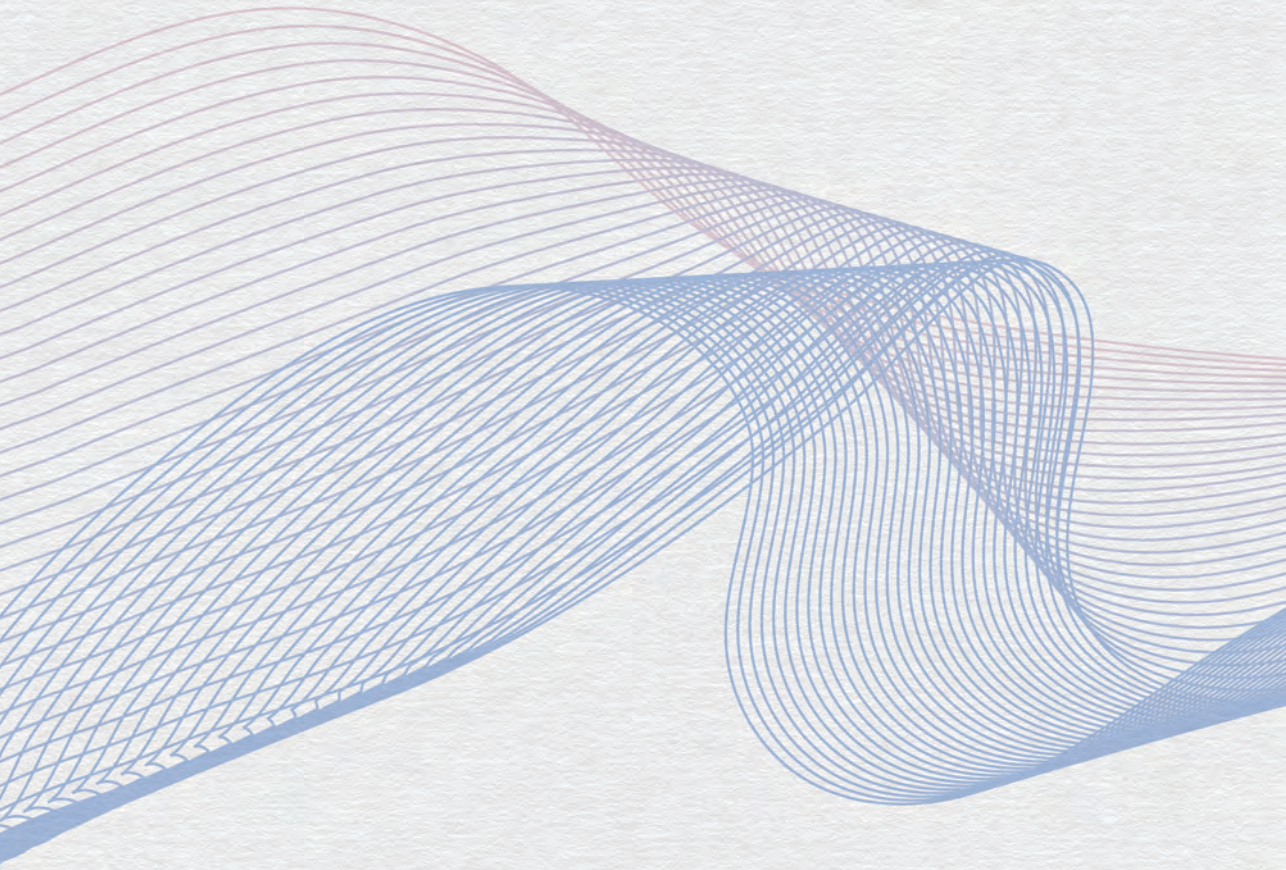
Item	Question
Confounding	5. Is there comparability of treatment groups, or are appropriate methods applied to correct for incomparability?
	5.1. RCT: Was the allocation sequence random?
	5.2. RCT: Was the allocation sequence concealed until participants were enrolled and assigned to interventions?
	5.3. RCT: Did baseline differences between intervention groups suggest a problem with the randomization process?
	5.4. Obs: Is there potential for confounding of the effect of the intervention in this study?
	5.5. Obs: Did the authors use an appropriate analysis method that controlled for all the important confounders?
	5.6. Obs: If 5.5. = Y or PY, were confounders that were controlled for measured adequately?
Missing data and selection bias	6. Were the patients included in the analysis representative of all patients included in the study and was the impact of missing data negligible?
	6.1. Were outcome data available for all, or nearly all, participants?
	6.2. Obs: Were intervention data available for all, or nearly all, participants?
	6.3. Obs: Were confounder data available for all, or nearly all, participants?
	6.4. If 6.1./6.2./6.3. = N or PN: were convincing arguments given for complete case analysis or were methods applied to address missing data?
	6.5. Was selection of participants into the study (or into the analysis) based on variables measured after the start of the intervention?
	6.6. Do start of follow-up and start of intervention coincide for all, or nearly all, participants?
Intervention status	7. Was intervention status correctly classified?
	7.1. Did the recorded intervention status correspond to the intervention actually received?
	7.2. Was there cross-over between interventions or non-adherence to the assigned intervention regimen that could have affected participants' outcomes?
	7.3. If 7.2. = Y or PY, was an appropriate analysis used to estimate the effect of starting and adhering to the intervention?
Outcome assessment	8. Was the outcome correctly measured?
	8.1. Was the outcome measurement a valid and reliable measurement of the outcome?
	8.2. Were outcome assessors aware of the intervention received by study participants?
	8.3. Were the methods of outcome assessment comparable across intervention groups?
Pre-specification of analysis	9. Were analyses prespecified and did the study adhere to the specified analysis plan?
	9.1. Was the analysis pre-specified, e.g., in a protocol?
	9.2. Are the reported results likely to be a selection of results of multiple analyses?
Red flags	Were there any aspects of the study or the report, **not covered by the other items**, that led to any doubt about the validity of the study? If yes, describe these in detail.

References

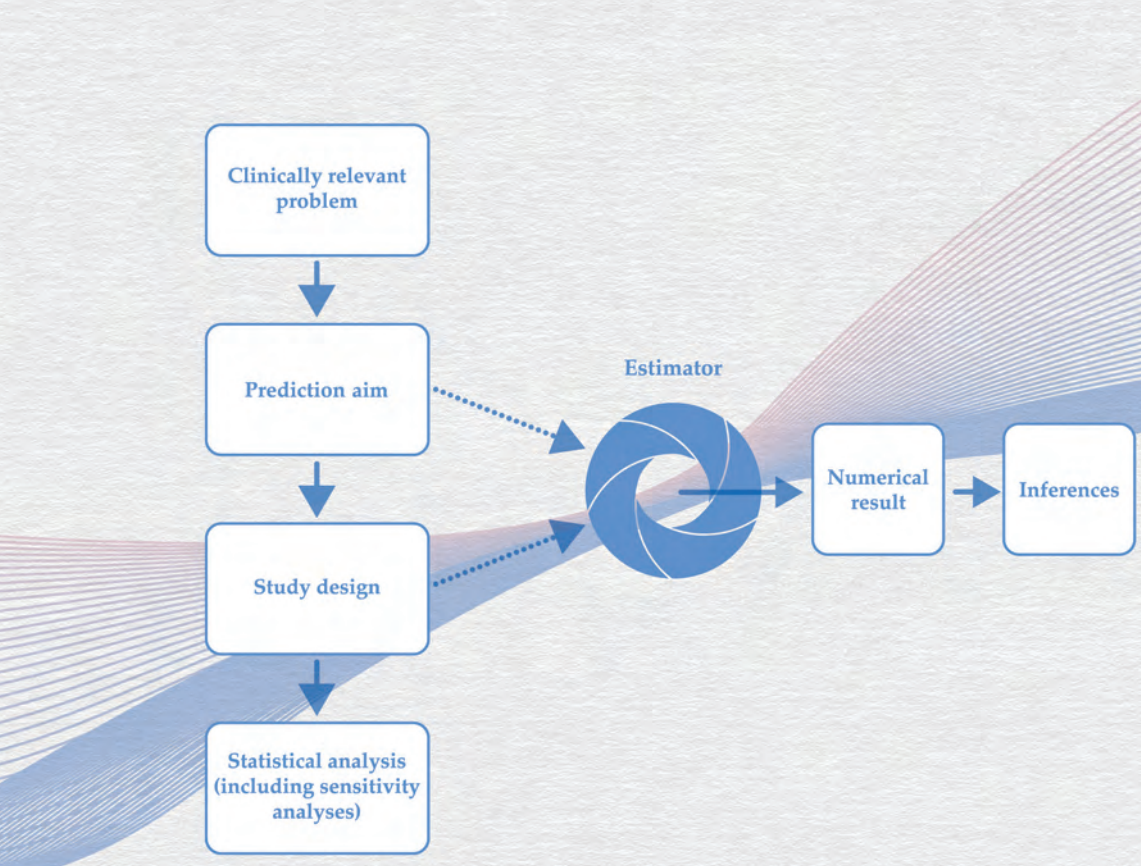
1. Robinson A, Johnson-Lynn S, Humphrey J, Haddad F. The challenges of translating the results of randomized controlled trials in orthopaedic surgery into clinical practice. *The Bone & Joint Journal*. 2019;101-B(2):121-123.
2. Axelrod D, Trask K, Buckley RE, Johal H. The Canadian Orthopaedic Trauma Society: lessons learned from 30 years of collaborative, high-impact research in fracture care. *The Bone & Joint Journal*. 2021;103(5):898-901.
3. Oberkofler CE, Hamming JF, Staiger RD, et al. Procedural surgical RCTs in daily practice: do surgeons adopt or is it just a waste of time? *Annals of Surgery*. 2019;270(5):727-734.
4. Chapman SJ, Aldaffaa M, Downey CL, Jayne DG. Research waste in surgical randomized controlled trials. *Journal of British Surgery*. 2019;106(11):1464-1471.
5. Houwert RM, van de Wall BJM, Groenwold RHH, Kruyt MC. A reaction to the editorial "Meta-Analyses and Systematic Reviews: JBJS Policy Revisited.". *The Journal of Bone and Joint Surgery*. 2021;103(10):849.
6. Beks RB, Bhashyam AR, Houwert RM, et al. When observational studies are as helpful as randomized trials: Examples from orthopedic trauma. *Journal of Trauma and Acute Care Surgery*. 2019;87(3):730-732.
7. Sanderson S, Tatt ID, Higgins J. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *International Journal of Epidemiology*. 2007;36(3):666-676.
8. Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar VS, Grimmer KA. A systematic review of the content of critical appraisal tools. *BMC Medical Research Methodology*. 2004;4(1):1-11.
9. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clinical Trials*. 1995;16(1):62-73.
10. D'Andrea E, Vinals L, Patorno E, et al. How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools. *British Medical Journal Open*. 2021;11(3):e043961.
11. Sterne JA, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *British Medical Journal*. 2019;366.
12. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *British Medical Journal*. 2016;355.
13. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. *ANZ Journal of Surgery*. 2003;73(9):712-716.
14. Jeyaraman MM, Rabbani R, Copstein L, et al. Methodologically rigorous risk of bias tools for nonrandomized studies had low reliability and high evaluator burden. *Journal of Clinical Epidemiology*. 2020;128:140-147.
15. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *Journal of Clinical Epidemiology*. 2020;126:37-44.
16. Minozzi S, Cinquini M, Gianola S, Castellini G, Gerardi C, Banzi R. Risk of bias in nonrandomized studies of interventions showed low inter-rater reliability and challenges in its application. *Journal of Clinical Epidemiology*. 2019;112:28-35.
17. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology*. 2011;64(4):395-400.
18. Hernán MA. Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*. 2016;26(10):674-680.

19. VanderWeele TJ. On well-defined hypothetical interventions in the potential outcomes framework. *Epidemiology (Cambridge, Mass).* 2018;29(4):e24.
20. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal*. 2008;337.
21. Fjalestad T, Hole MØ, Hovden IAH, Blücher J, Strømsøe K. Surgical treatment with an angular stable plate for complex displaced proximal humeral fractures in elderly patients: a randomized controlled trial. *Journal of Orthopaedic Trauma*. 2012;26(2):98-106.
22. Knobe M, Drescher W, Heussen N, Sellei RM, Pape H-C. Is helical blade nailing superior to locked minimally invasive plating in unstable pertrochanteric fractures? *Clinical Orthopaedics and Related Research*. 2012;470(8):2302-2312.
23. Constant C, Murley A. A clinical method of functional assessment of the shoulder. *Clinical Orthopaedics and Related Research*. 1987(214):160-164.
24. Rangan A, Handoll H, Brealey S, et al. Surgical vs nonsurgical treatment of adults with displaced fractures of the proximal humerus: the PROFHER randomized clinical trial. *Journal of the American Medical Association*. 2015;313(10):1037-1047.
25. Altman DG, Schulz KF. Concealing treatment allocation in randomised trials. *British Medical Journal*. 2001;323(7310):446-447.
26. Hernán MA, Robins JM. *Causal inference: what if.* In: Boca Raton: Chapman & Hall/CRC; 2020.
27. Beks RB, Reetz D, de Jong MB, et al. Rib fixation versus non-operative treatment for flail chest and multiple rib fractures after blunt thoracic trauma: a multicenter cohort study. *European Journal of Trauma and Emergency Surgery*. 2019;45(4):655-663.
28. Houwert RM, Beks RB, Dijkgraaf MG, Roes KC, Öner FC, Hietbrink F, Leenen LP, Groenwold RHH. Study methodology in trauma care: towards question-based study designs. *European Journal of Trauma and Emergency Surgery*. 2021;47(2):479-84.
29. Jenkinson RJ, Kiss A, Johnson S, Stephen DJ, Kreder HJ. Delayed wound closure increases deep-infection rate associated with lower-grade open fractures: a propensity-matched cohort study. *Journal of Bone and Joint Surgery*. 2014;96(5):380-386.
30. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass)*. 2009;20(4):488.
31. VanderWeele TJ. On the relative nature of overadjustment and unnecessary adjustment. *Epidemiology*. 2009;20(4):496-499.
32. Groenwold RH, Klungel OH, Grobbee DE, Hoes AW. Selection of confounding variables should not be based on observed associations with exposure. *European Journal of Epidemiology*. 2011;26(8):589.
33. Sun G-W, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology*. 1996;49(8):907-916.
34. Heinze G, Dunkler D. Five myths about variable selection. *Transplant International*. 2017;30(1):6-10.
35. Altman DG, Royston P. The cost of dichotomising continuous variables. *British Medical Journal*. 2006;332(7549):1080.
36. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*. 2013;177(4):292-298.
37. Portinari M, Bianchi L, De Troia A, et al. Non-traumatic emergency abdominal surgery in nonagenarian patients: a retrospective study. *European Journal of Trauma and Emergency Surgery*. 2021:1-12.
38. Lee KJ, Tilling K, Cornish RP, et al. Framework for the Treatment And Reporting of Missing data in Observational Studies: The TARMOS framework. *arXiv preprint arXiv:200414066*. 2020.
39. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*. 2009;338.
40. Carpenter JR, Smuk M. Missing data: A statistical framework for practice. *Biometrical Journal*. 2021;63(5):915-947.
41. Groenwold RH, Dekkers OM. Missing data: the impact of what is not there. *European Journal of Endocrinology*. 2020;183(4):E7-E9.

42. Klei DS, Öner FC, Leenen LP, van Wessem KJ. No need for sternal fixation in traumatic sternovertebral fractures: outcomes of a 10-year retrospective cohort study. *Global Spine Journal*.2192568220902413.
43. Van der Meijden OA, Houwert RM, Hulsmans M, et al. Operative treatment of dislocated midshaft clavicular fractures: Plate or intramedullary nail fixation?: A randomized controlled trial. *Journal of Bone and Joint Surgery*. 2015;97(8):613-619.
44. Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable? *Injury*. 2011;42(3):236-240.
45. Banierink H, Reininga I, Heineman E, Wendt K, Ten Duis K, IJpma F. Long-term physical functioning and quality of life after pelvic ring injuries. *Archives of Orthopaedic and Trauma Surgery*. 2019;139(9):1225-1233.
46. Ochen Y, Frima H, Houwert RM, et al. Surgical treatment of Neer type II and type V lateral clavicular fractures: comparison of hook plate versus superior plate with lateral extension: a retrospective cohort study. *European Journal of Orthopaedic Surgery & Traumatology*. 2019;29(5):989-997.
47. Nauth A, Creek AT, Zellar A, Lawendy AR, Dowrick A, Gupta A, Dadi A, van Kampen A, Yee A, de Vries AC, van Otterloo AD. Fracture fixation in the operative management of hip fractures (FAITH): an international, multicentre, randomised controlled trial. *The Lancet*. 2017;389(10078):1519-27.
48. Groenwold RH, Goeman JJ, Le Cessie S, Dekkers OM. Multiple testing: when is many too much? *European Journal of Endocrinology*. 2021;184(3):E11-E14.
49. Gamble C, Krishan A, Stocken D, et al. Guidelines for the content of statistical analysis plans in clinical trials. *Journal of the American Medical Association*. 2017;318(23):2337-2343.
50. Chan A-W, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Annals of Internal Medicine*. 2013;158(3):200-207.
51. Loder E, Groves T, MacAuley D. Registration of observational studies. *British Medical Journal*. 2010;340:c950.
52. Smeeing DPJ, Houwert RM, Briet JP, et al. Weight-bearing or non-weight-bearing after surgical treatment of ankle fractures: a multicenter randomized controlled trial. *European Journal of Trauma and Emergency Surgery*. 2020;46(1):121-130.
53. Briet JP, Houwert RM, Smeeing DP, et al. Weight bearing or non-weight bearing after surgically fixed ankle fractures, the WOW! Study: study protocol for a randomized controlled trial. *Trials*. 2015;16(1):1-8.
54. Barkun JS, Aronson JK, Feldman LS, Maddern GJ, Strasberg SM, Collaboration B. Evaluation and stages of surgical innovations. *The Lancet*. 2009;374(9695):1089-1096.
55. Ergina PL, Cook JA, Blazeby JM, et al. Challenges in evaluating surgical innovation. *The Lancet*. 2009;374(9695):1097-1104.
56. McCulloch P, Altman DG, Campbell WB, et al. No surgical innovation without evaluation: the IDEAL recommendations. *The Lancet*. 2009;374(9695):1105-1112.
57. Khachane A, Philippou Y, Hirst A, McCulloch P. Appraising the uptake and use of the IDEAL framework and recommendations: a review of the literature. *International Journal of Surgery*. 2018;57:84-90.
58. Bilbro NA, Hirst A, Paez A, et al. The ideal reporting guidelines: a Delphi consensus statement stage specific recommendations for reporting the evaluation of surgical innovation. *Annals of Surgery*. 2021;273(1):82-85.
59. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*. 2008;336(7650):924-926.
60. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *Journal of Clinical Epidemiology*. 2011;64(4):380-382.
61. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. *British Medical Journal*. 2002;324(7351):1448-1451.
62. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *British Medical Journal Evidence-Based Medicine*. 2016;21(4):125-127.
63. Meakins JL. Evidence-based practice: new techniques and technology. *Canadian Journal of Surgery*. 2001;44(4):247-249.



6


Clinically relevant problem
Prediction aim
Study design
Statistical analysis (including sensitivity analyses)
Estimator
Numerical result
Inferences

Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective

It is widely acknowledged that the predictive performance of clinical prediction models should be studied in patients that were not part of the data in which the model was derived. Out-of-sample performance can be hampered when predictors are measured differently at derivation and external validation. This may occur, for instance, when predictors are measured using different measurement protocols or when tests are produced by different manufacturers. Although such heterogeneity in predictor measurement between derivation and validation data is common, the impact on the out-of-sample performance is not well studied. Using analytical and simulation approaches, we examined out-of-sample performance of prediction models under various scenarios of heterogeneous predictor measurement. These scenarios were defined and clarified using an established taxonomy of measurement error models. The results of our simulations indicate that predictor measurement heterogeneity can induce miscalibration of prediction and affects discrimination and overall predictive accuracy, to extents that the prediction model may no longer be considered clinically useful. The measurement error taxonomy was found to be helpful in identifying and predicting effects of heterogeneous predictor measurements between settings of prediction model derivation and validation. Our work indicates that homogeneity of measurement strategies across settings is of paramount importance in prediction research.

This chapter was based on: Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. *Statistics in Medicine*. 2019;38(18):3444-3459.

1 | Background

Prediction models have an important role in contemporary medicine by providing probabilistic predictions of diagnosis or prognosis[1]. Prediction models need to provide accurate and reliable predictions for patients that were not part of the dataset in which the model was derived (i.e., derivation set)[2]. The ability of a prediction model to predict in future patients (i.e., out-of-sample) can be evaluated in an external validation study. While out-of-sample predictive performance is in general expected to be lower than performance estimated at derivation[1], large discrepancies are often contributed to suboptimal modelling strategies in the derivation of the model[3-5] and differences between patient characteristics in derivation and validation samples[6,7].

Another potential source of limited out-of-sample performance is when predictors are measured differently at derivation than at (external) validation. This may occur, for instance, when predictors are categorized using different cut-off values or when predictors are based on diagnostic tests that were produced by different manufacturers (see Table 1 for examples). Although some studies have mentioned that such heterogeneity in predictor measurements might hamper out-of-sample model performance (e.g.,[8,9]), effects of measurement heterogeneity in prediction studies have received little attention. Particularly, its impact on predictive performance has not been formally quantified.

In this study, we investigate the out-of-sample performance of a clinical prediction model in situations where predictor measurement strategies at the model derivation stage differed from measurement strategies at the model validation stage. The different scenarios of heterogeneous predictor measurement were defined using a well-known taxonomy of measurement error models, described by e.g., Keogh and White[16]. We varied the degree of measurement error in the derivation data and validation data to recreate qualitative differences in the predictor measurement structures across settings. Hence, the measurement error perspective serves as a framework to define predictor measurement heterogeneity. We focus on logistic regression, since this model is widely applied in clinical prediction research[17].

This paper is structured as follows. In Section 2, we define the measurement error models used to describe scenarios of measurement heterogeneity. In Section 3, we derive analytical expressions to identify and predict effects of measurement error on in-sample predictive performance. In Section 4, we illustrate the effects of measurement heterogeneity across settings on predictive performance in large sample simulations

and contrast these to the impact of measurement error within the derivation setting. In Section 5, we present an extensive set of Monte Carlo simulations in finite samples to examine the impact of measurement heterogeneity on out-of-sample predictive performance. We end with discussing the implications of our findings in Section 6.

Table 1 Possible sources of measurement heterogeneity in measurements of predictors, illustrated by examples from previously published prediction studies

Type of predictor	Examples of predictors	Examples of measurement heterogeneity
Anthropometric measurements	Height Weight Body circumference	Guidelines on imaging decisions in osteoporosis care are established using standardized measurements of height, while in clinical practice height is measured using non-standardized techniques or self-reported values[10].
Physiological measurements	Blood pressure Serum cholesterol HbA1c Fasting glucose	In scientific studies, blood pressure is often measured by the average of multiple measurements performed under standardized conditions, while blood pressure measurements in practice deviate from protocol guidelines in various ways due to variability in available time and devices[11].
Diagnosis	Previous / current disease	The diagnosis 'hypertension' can be defined as a blood pressure of ≥ 140/90 mmHg (without use of anti-hypertensive therapy) or as the use of anti-hypertensive drugs[12].
Treatment/ Exposure status	Type of drug used Smoking status Dietary intake	The cut-off value for an 'increased length of stay in the hospital' to predict unplanned readmission may depend on the country in which the model is evaluated[13].
Imaging	Presence or size of tissue on ultrasound, MRI, CT, of FDG PET scans	In scientific studies, review of FDG PET scans may be protocolized or performed by a single experienced nuclear medicine physician, blinded to patient outcome[14]. In routine practice, FDG PET scans may be reviewed under various systematics or by a multi-disciplinary team[15].

2 | Expressing measurement heterogeneity in terms of measurement error models

Consider a random sample of N independent individuals $i = 1, \ldots, N$. Let Y be a binary response variable with values $y_i \in \{0, 1\}$. We define a logistic regression model for estimating the probability that $Y = 1$ given values of a set of P continuous predictor

variables, $\boldsymbol{X} = \{X_1, \ldots, X_p\}$. The probability of observing an event ($Y = 1$) given the predictors, $\pi_i = P(Y_i = 1|\boldsymbol{X}_i)$, is defined as

$$\pi_i = \frac{1}{1 + \exp(-(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i))},$$

where α is an intercept (scalar), $\boldsymbol{\beta}$ is a P-dimensional vector of regression coefficients.

For simplicity of presentation, we consider a single vector $X \subset \boldsymbol{X}$. To distinguish different measurements of the same predictor, we denote an exact measurement of the predictor (e.g., bodyweight measured on a scale) by X and a pragmatic measurement (e.g., self-reported weight) by W. In most measurement error literature, X denotes an error-free true value and W denotes an observed error-prone version of X[18]. However, for prediction purposes, it is hardly ever feasible (or even undesirable) to obtain error-free measurements in clinical practice, and hence we use the terms exact measurement for X and pragmatic measurement for W. The connection between X and W can be formally defined using measurement error models. We define a general model of measurement heterogeneity for continuous predictors in line with existing measurement error literature[16,18]. Assuming that the relation between X and W is linear and additive, the association between W and X can be described as

$$\begin{aligned} \mathbb{E}(W|Y = y) &= \psi_{Y=y} + \theta_{Y=y}\mathbb{E}(X) + \epsilon_{Y=y}, \\ \mathrm{Var}(W|Y = y) &= \theta^2_{Y=y}\sigma^2_X + \sigma^2_{\epsilon_{Y=y}}, \end{aligned} \tag{1}$$

where $\epsilon_{Y=y} \sim \mathcal{N}(0, \sigma^2_{\epsilon_{Y=y}})$ and all parameters may depend on the value of Y, indicating that measurements can differ between individuals in which the outcome is observed (cases) and individuals in which the outcome is not observed (non-cases). The parameter ψ reflects the mean difference between X and $W \mid Y = y$, θ indicates the linear association between X and $W \mid Y = y$, and σ^2_ϵ reflects variance introduced by random deviations in the measurement process, where a larger σ^2_ϵ indicates that the measurement W is less precise. The term *measurement error* applies to situations where both an exact measurement and a pragmatic measurement of a predictor are available within a setting (e.g., the derivation set), and thus where the parameters ψ, θ and σ^2_ϵ define the degree of measurement error in W with respect to X. The term *measurement heterogeneity* refers to situations where the same predictor is measured heterogeneously across settings of derivation and validation. The most precise measurement (whether available at derivation or validation) corresponds to X and the parameters ψ, θ and σ^2_ϵ define the degree of heterogeneity between X and W. We now consider three types of measurement error models that are particular forms of

Equation (1), based on which we specify both within-sample measurement error and measurement heterogeneity across settings.

2.1 | Random measurement error model

Under $\psi = 0$ and $\theta = 1$, Equation (1) reduces to the following model:

$$\mathbb{E}(W) = \mathbb{E}(X) + \epsilon, \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$is independent of X and Y. This is referred to as the random or classical measurement error model[16,18]. W is a mean-unbiased measurement of X, since $\mathbb{E}(W|Y) = \mathbb{E}(W) = \mathbb{E}(X)$. An example of a predictor measurement corresponding to the random measurement error model is reading body weight from the same scale. Each reading, the value may deviate slightly upwards or downwards, resulting in random deviations. Variation in the size of these deviations across settings due to precision of the available scales is an example of random measurement heterogeneity.

2.2 | Systematic measurement error model

When $\psi \neq 1$ and/or $\theta \neq 1$, yet when ψ and θ have the same values for cases and non-cases, predictor measurements correspond to a systematic measurement error model[16]. The systematic measurement error model is defined as

$$\mathbb{E}(W) = \psi + \theta\mathbb{E}(X) + \epsilon, \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is independent of X and Y. It follows that W is no longer a mean-unbiased measurement of X ($\mathbb{E}(W) \neq \mathbb{E}(X)$). Systematic measurement heterogeneity may occur, for example, when a blood glucose monitor is replaced by a monitor from a different manufacturer that is calibrated differently. The switch in measurement instrument may introduce a shift by a constant in the measured predictor values, i.e., a change in ψ (additive systematic measurement error). Furthermore, observed values may depend on the actual value of a predictor, where θ represents linear dependencies between X and W. For instance, values of self-reported weight may be underreported, especially by individuals with a higher actual weight, i.e., $\theta < 1$ (multiplicative systematic measurement error). The size of ψ and θ can differ across settings, for example when weight is measured using a scale in one setting (e.g., θ might be close to 1) and as a self-reported value in another setting (e.g., θ might deviate from 1), which would result in systematic measurement heterogeneity.

2.3| Differential measurement error model

In case measurement procedures differ between cases and non-cases, i.e., when $\psi_1 \neq \psi_0$ and/or $\theta_1 \neq \theta_0$ and/or $\sigma^2_{\epsilon 1} \neq \sigma^2_{\epsilon 0}$, the measurements can be described by Equation (1) above, also referred to as differential measurement error[16]. Differential measurement of predictors is conceivable in settings where assessment of predictors is done in an unblinded fashion, such as case-control studies[19]. For example, when patient history is collected after observing the outcome event, cases may be more likely to recall health information prior to the outcome event than non-cases, also known as recall bias[20]. This may for example lead to over-reporting in cases, i.e., $\psi_1 > \psi_0$, a stronger association between reported and actual predictor values, i.e., $\theta_1 > \theta_0$, or more precise predictor measurements, i.e., $\sigma^2_{\epsilon_1} < \sigma^2_{\epsilon_0}$, in cases than in non-cases. Prospective differential measurement error may occur when a prediction model influences the way that predictors are measured in clinical practice. After clinical uptake of a prediction model, physicians may measure predictors differently in patients in whom they suspect the outcome of interest (potential future cases), guided by the knowledge that these particular predictors are of importance. For example, in these patients, body weight may be measured using a scale, whereas the prediction model may have been derived from self-reported measurements of body weight, introducing a difference between measurement procedures of (potential) cases and non-cases (i.e., differential measurement error), as well as a difference in measurement strategy between derivation and application setting (i.e., differential measurement heterogeneity).

3 | Predictive performance under within-sample measurement error

In this section, we define analytical expressions that indicate how substituting an exact predictor measurement, X, with a pragmatic predictor measurement, W, affects apparent predictive performance in the situation where both measurements X and W are available in the derivation sample of a prediction model. For brevity, we will evaluate a single-predictor model. Expressions of in-sample predictive performance under random measurement error were previously derived by Khudyakov and colleagues for a probit prediction model[21]. The current paper extends these expressions to a logistic regression model. We measure predictive performance by the concordance-statistic (c-statistic) and Brier score, measuring discrimination and overall accuracy, respectively. Effects on calibration will be evaluated in the next sections. We will discuss expressions in terms of sample realizations, that is, realizations y_i, x_i and w_i. In the following, let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i|y_i)$ and s^2_x denote the sample mean and variance of x,

let $\bar{w} = \frac{1}{n}\sum_{i=1}^{n}(w_i|y_i)$ and s_w^2 denote the sample mean and variance of w, and let n_1 and n_0 denote the number of cases and non-cases in the sample, respectively.

3.1| C-statistic

To examine the discriminatory performance, we make use of the c-statistic, a rank-order statistic that typically ranges from 0.5 (no discrimination) to 1 (perfect discrimination) and is equal to the area under the receiver operating characteristic (ROC) curve for a binary outcome[17]. Consider a data-generating model relating response variable Y to X by a logit link function, where $X|Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ (binormality). Let $\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i|y_i = 1)$ denote the sample mean of x for cases, let $\bar{x}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0}(x_i|y_i = 0)$ denote the sample mean of x for non-cases, and let $s_{x1}^2 + s_{x0}^2$ denote the total variance of x. Let Φ denote the cumulative distribution function of the standard normal distribution. Following Austin and Steyerberg[22], the c-statistic is approximated by

$$AUC_x = \Phi\left(\frac{\bar{x}_1 - \bar{x}_0}{\sqrt{s_{x1}^2 + s_{x0}^2}}\right).$$

Alternatively, for w, let $\bar{w}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}(w_i|y_i = 1)$ and $\bar{w}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0}(w_i|y_i = 0)$ denote the sample means of w for cases and non-cases, respectively, and let $s_{w1}^2 + s_{w0}^2$ denote the total variance of w. The c-statistic of a binary logistic regression model of the predictor w is then given by:

$$AUC_w = \Phi\left(\frac{\bar{w}_1 - \bar{w}_0}{\sqrt{s_{w1}^2 + s_{w0}^2}}\right). \tag{4}$$

Under the general measurement error model (Equation (1)),

$$\begin{aligned}
\bar{w}_0 &= \psi_0 + \theta_0\bar{x}_0,\\
\bar{w}_1 &= \psi_1 + \theta_1\bar{x}_1,\\
s_{w0}^2 &= s_{x0}^2\theta_0^2 + s_{\epsilon_0}^2,\\
s_{w1}^2 &= s_{x1}^2\theta_1^2 + s_{\epsilon_1}^2.
\end{aligned}$$

The impact of measurement error on the c-statistic can now be expressed as

$$\begin{aligned}
\Delta AUC &= AUC_w - AUC_x\\
&= \Phi\left(\frac{(\psi_1 + \theta_1\bar{x}_1) - (\psi_0 + \theta_0\bar{x}_0)}{\sqrt{s_{x1}^2\theta_1^2 + s_{\epsilon_1}^2 + s_{x0}^2\theta_0^2 + s_{\epsilon_0}^2}}\right) - \Phi\left(\frac{\bar{x}_1 - \bar{x}_0}{\sqrt{s_{x1}^2 + s_{x0}^2}}\right),
\end{aligned} \tag{5}$$

where a $\Delta AUC < 0$ indicates that the model has less discriminatory power when w is used instead of x. Equations (4) and (5) indicate that the expected impact of substituting x by w in prediction model development has the following consequences. In case of random measurement error in w, it can be expected that the model fitted on w has a lower c-statistic and $\Delta AUC < 0$. In case of systematic measurement error in w, the c-statistic is not affected beyond random measurement error. Differential measurement error can affect model discrimination in both directions. For example, when observed measurements w are systematically shifted further from x in cases, i.e., when $\psi_1 > \psi_0$ and $\theta_1 = \theta_0 = 1$, and when the difference in mean predictor values between cases and non-cases in x is positive, i.e., $\bar{x}_1 > \bar{x}_0$ and $AUC_X > 0.5$, the mean difference in predictor values between cases and non-cases, $\bar{w}_1$ - $\bar{w}_0$, increases, enlarging the discriminatory power of the model, i.e., $\Delta AUC > 0$. Additional random measurement error affects the c-statistic irrespective of whether the error is differential or not.

3.2 | Brier score

As a measure of overall predictive accuracy, we evaluate the Brier score, which is a proper scoring rule that indicates the distance between predicted and observed outcomes. The Brier score is calculated by[23]

$$BS(x) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\pi}(x_i))^2, \tag{6}$$

Where $\hat{\pi}(x_i) = (1 + \exp(-(\hat{\alpha}_x + \hat{\beta}_x x_i)))^{-1}$ and a lower Brier score indicates higher accuracy of predictions. Following Blattenberger and Lad[24] and Spiegelhalter[25], the Brier score can be decomposed into

$$BS(x) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\pi}(x_i))(1 - 2\hat{\pi}(x_i)) + \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}(x_i)(1 - \hat{\pi}(x_i)), \tag{7}$$

resulting in a calibration component, $(y_i - \hat{\pi}(x_i))(1 - 2\hat{\pi}(x_i))$, and a refinement component, $\hat{\pi}(x_i)(1 - \hat{\pi}(x_i))$. As Spiegelhalter already noted[25], the calibration component has an expectation of 0 under the null hypothesis of perfect calibration, that is $\mathbb{E}_0(Y_i) = \hat{\pi}(x_i)$, and the expected Brier score can be expressed by the refinement term in Equation (7), that is $\mathbb{E}_0(BS(x)) = \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}(x_i)(1 - \hat{\pi}(x_i))$. Consequently, the impact of within-sample measurement error on the Brier score of a maximum likelihood model in the derivation set can be expressed as

$$\mathbb{E}_0(\Delta BS) = \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}(w_i)(1-\hat{\pi}(w_i)) - \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}(x_i)(1-\hat{\pi}(x_i)), \qquad (8)$$

where

$$\hat{\pi}(w_i) = \frac{1}{1+\exp(-(\hat{\alpha}_w + \hat{\beta}_w(\psi_{Y=y} + x_i\theta_{Y=y} + \epsilon_{Y=y}))},$$

where a $\mathbb{E}_0(\Delta BS) > 0$ indicates that substituting x with w yields less accurate predictions. Realistically, however, a model is hardly ever perfectly calibrated (see[26] for an in-depth discussion of levels of calibration of prediction models). A maximum likelihood estimate of a logistic regression model attains "weak calibration" in its derivation sample by definition, meaning that no systematic overfitting or underfitting and/or overestimation or underestimation of risks occurs. In the remaining of this paper, we use the term "calibration" instead of "weak calibration" and use the term "Brier score" to refer to the decomposed empirical Brier score in Equation (7).
Expression (8) indicates that substituting x with w in a perfectly specified model has the following consequences. When the association between w and outcome y is weaker than the association between x and y, a prediction model based on w provides less extreme predicted probabilities. This results in a larger refinement term for w, i.e., $\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}(w_i)(1-\hat{\pi}(w_i))$ is larger, and in a positive $\mathbb{E}_0(\Delta BS)$and hence lower accuracy.

4 | Measurement error versus measurement heterogeneity

The expressions of predictive performance under measurement error indicate that more erroneous predictor measurements lead to less apparent discriminatory power and accuracy. However, these results cannot be generalized directly to effects of measurement error on out-of-sample performance of prediction models. We use the measurement error model taxonomy to explore how heterogeneity in measurement structures affects out-of-sample performance. Rather than distinguishing error-free and error-prone predictor measurements, the measurement error models now express deviations from homogeneity of measurements across settings.

A direct comparison of effects of measurement error and effects of measurement heterogeneity on predictive performance can be found in Figures 1 and 2, which illustrate large-sample (N = 1,000,000) properties of predictive performance measures. Effects of measurement error are illustrated by comparing in-sample predictive performance measures of a prediction model that is first estimated based on x and subsequently estimated based on w, where the latter contains increasing measurement

error. Effects of measurement heterogeneity are illustrated by comparing out-of-sample predictive performance measures of a prediction model that is transported across settings with different predictor measurement structures. We explored three settings: (i) x is available at derivation and w is available at validation, (ii) w is available at both derivation and validation, and (iii) w is available at derivation and x is available at validation. In other words, this section illustrates the impact of measurement error and

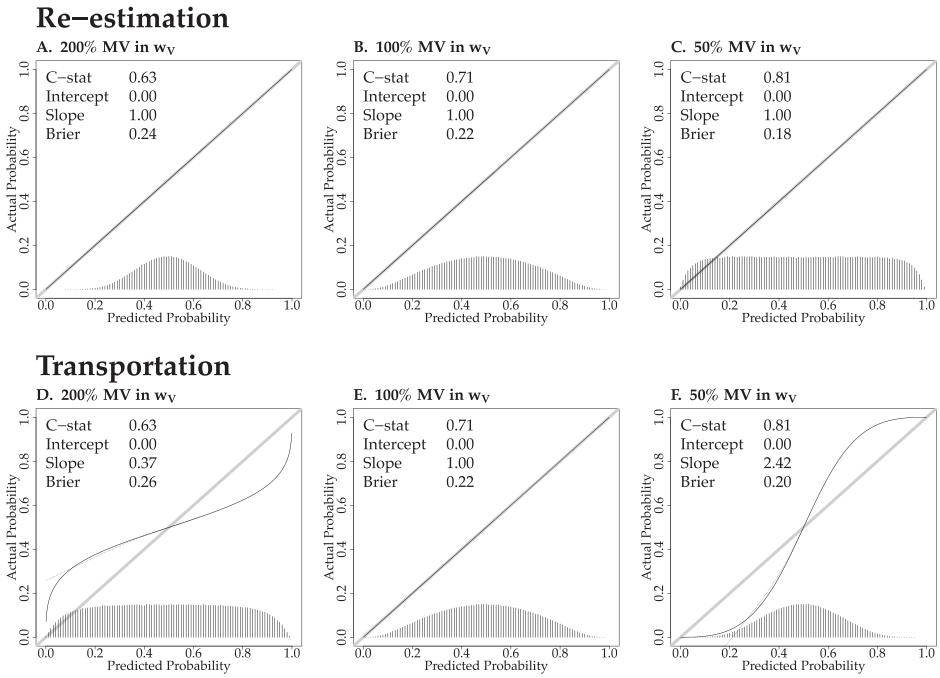


Figure 1. Measures of predictive performance under predictor measurement error and predictor measurement heterogeneity. The data-generating mechanism corresponded perfectly to the estimated logistic regression model. The top rows show calibration plots of a single-predictor model that is fitted using predictor measurement x and validated by re-estimating the model on the same data using w, i.e., under predictor measurement error. The bottom rows show situations where the same model is transported from derivation to validation setting, specifically, the model (D) is derived using x and validated using w, (E) is derived and validated using w, and (F) is derived using w and validated using x, i.e., under predictor measurement heterogeneity. The calibration plots show the calibration slope (black line) and predicted probability frequencies (bottom-histograms) for situations in which the predictor measurement variance at validation equals 200% (A, D), 100% (B, E), or 50% (C, F) of the predictor measurement variance at derivation. The val.prob function from the `rms` package was used to compute the simulation outcome measures and to generate the calibration plots[27], where we edited the legend format settings in the plot to improve readability. MV = measurement variance of the predictor measurement used for model validation relative to the predictor measurement used for derivation.

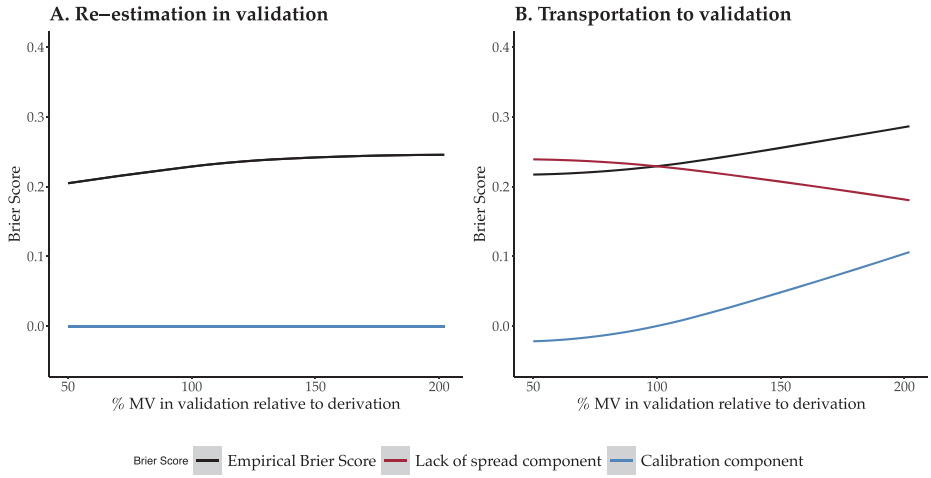


Figure 2. Decomposed Brier score under predictor measurement error and predictor measurement heterogeneity. The data-generating mechanism corresponded perfectly to the estimated logistic regression model. The plot displays the large sample properties of the components of the Brier score (Equation 7) under increasing random predictor measurement variance at validation, corresponding to the random measurement error model (Equation 2). The left panel shows the Brier score for a single-predictor logistic regression model that is fitted using predictor measurement x and validated by re-estimating the model on the same data using w, i.e., under predictor measurement error. The right panel shows transportation from w at derivation to x at validation (up to %MV = 100) and transportation from x at derivation to w at validation (from %MV = 100 onwards) i.e., under predictor measurement heterogeneity. MV = measurement variance of the predictor measurement used for model validation relative to the predictor measurement used for derivation.

measurement heterogeneity as an isolated factor by evaluating the same population at both derivation and validation, and only varying the predictor measurement structures. For the purpose of demonstration, we focus on random measurement error and -heterogeneity and provide further analyses in the next section.

In situations of within-sample measurement error, i.e., in the re-estimated model, all calibration plots showed a calibration slope equal to $b = 1$, indicating perfect apparent calibration (Figure 1A-C). The apparent c-statistic and Brier score improved with decreasing random measurement error. In case of measurement heterogeneity across samples, i.e., in the transported model, similar changes in the c-statistic and Brier score were found. However, heterogeneous measurements led to a calibration slope $b \neq 1$, indicating that predictions were no longer valid (Figure 1D and 1F). When measurements at validation were less precise than at derivation, the calibration slope was $b < 1$, similar to statistical overfitting. When measurements at validation were

more precise than at derivation, the calibration slope was $b > 1$, similar to statistical underfitting. More elaborate illustrations of the impact of measurement heterogeneity in large sample simulations, including effects of systematic and differential measurement heterogeneity, can be found in the Online Supplementary File 1.

Although the total Brier score did not differ substantially between the re-estimated and transported model, the examination of the large sample properties of the decomposed Brier score (Equation (7)) indicated differences in the components between the procedures (Figure 2). In the re-estimated model, the calibration term equalled zero, and the total Brier score equalled the refinement term (Figure 2A). The Brier score increased with increasing random measurement error, indicating that accuracy decreased. In the transported model, changes in the refinement term were counterbalanced by changes in the calibration term. For example, when measurements at validation were less precise than at derivation, the spread in predicted probabilities increased (refinement term in Figure 2B decreased). A decrease in the refinement term under perfect calibration would indicate that overall accuracy of the model is improving, as predicted probabilities are closer to 0 or 1. However, in the transported model this improvement was counterbalanced by a calibration term larger than zero, which indicates that predicted probabilities were too extreme compared to observed probabilities (Figure 2B).

Figures 1 and 2 illustrate that miscalibration is not introduced by measurement error per se but rather by measurement heterogeneity across settings of derivation and validation. The discrepancy in calibration between model re-estimation and model transportation can be reduced to differences in the linear predictors of the recalibration models. In case of model re-estimation, the linear predictor is expressed by

$$\mathrm{lp}_{re-est} = \hat{\alpha}_{w(V)} + \hat{\beta}_{w(V)} w_{iV}, \tag{9}$$

indicating that the parameters $\hat{\alpha}_{w(V)}$ and $\hat{\beta}_{w(V)}$ are estimated using the predictor values measured by strategy w in the validation data. In the more realistic validation procedure in which the model is transported over different predictor measurement procedures, the linear predictor is expressed by

$$\mathrm{lp}_{transp} = \hat{\alpha}_{x(D)} + \hat{\beta}_{x(D)} w_{iV}, \tag{10}$$

meaning that regression coefficients are estimated based on x_{iD} and that the model is validated using w_{iV}. This distinction in recalibration models sheds a different light on previous research into effects of measurement error on predictive performance. Khudyakov and colleagues derived analytically that calibration in a derivation sample is not affected

by measurement error[21]. Since their findings are based on the assumption that the linear predictor is defined as in Equation (9), previous results on the impact of measurement error on predictive performance can be interpreted as effects on in-sample predictive performance[21,29].

5 | Predictive performance under measurement heterogeneity across settings

5.1| Simulation methods

General patterns of predictive performance under measurement heterogeneity were examined in a set of Monte Carlo simulations in finite samples to evaluate their behavior under sampling variability. Simulations were performed in R version 3.3.1[30], and our code is accessible online (see https://github.com/KLuijken/Prediction_Measurement_Heterogeneity_Predictor). We studied the predictive performance of a single- and a two-predictor binary logistic regression model. For the latter, we evaluated situations in which both predictors were measured heterogeneously across settings as well as situations in which one of the predictors was measured similar over settings. The data for the single-predictor model were generated from

$$\begin{aligned} &\text{logit}(Y) = \log(4)X, \\ &\text{where } X \sim \mathcal{N}(0,1). \end{aligned}$$

The data for the two-predictor models were generated from

$$\begin{aligned} &\text{logit}(Y) = \boldsymbol{\beta}^T \boldsymbol{X}, \\ &\text{where } \boldsymbol{X} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{X1X2} \\ \rho_{X1X2} & 1 \end{pmatrix} \right). \end{aligned}$$

The correlation between predictors, ρ_{X1X2}, varied with 0, 0.5 and 0.9. Both the β-parameters in the two-predictor models have value 2.3 in case $\rho_{X1X2} = 0$ or $\rho_{X1X2} = 0.5$, and have value 2.1 in case $\rho_{X1X2} = 0.9$. We varied the values of the regression coefficients in order to keep the c-statistic of the data-generating models at an approximate value of 0.80 and hence to compare predictive performance over models[22]. We recreated different measurement procedures of the predictors using different specifications of the general measurement error model (Equation (1)). In the derivation sample, measurements corresponded to the random measurement error model (Equation (2)), while in validation various measurement structures were recreated (see Table 2 for values of input parameters). All measurements contained at least some erroneous measurement variance to generate realistic scenarios.

Table 2 Input parameters for finite sample simulations. Full-factorial simulations for the parameters ψ, θ and σ_ϵ resulted in 54 scenarios for the single-predictor model, and 162 scenarios in both the two-predictor model with and the model without a predictor that was measured homogeneously across settings. An additional 54 scenarios of differential measurement error in the single-predictor model were evaluated, resulting in a total of 432 scenarios.

	Simulation parameter	Factor values
Derivation	ψ_D	0
	θ_D	1.0
	$\sigma_{\epsilon(D)}$	0.5, 1.0, 2.0
Validation	ψ_V	0, 0.25
	θ_V	0.5, 1.0, 2.0
	$\sigma_{\epsilon(V)}$	0.5, 1.0, 2.0

In total, 432 scenarios were evaluated. For each scenario, a derivation sample (n = 2,000) and a validation sample (n = 2,000) were generated. We did not consider smaller sample sizes, since predictive performance measures are sensitive to statistical overfitting, which would complicate the interpretation of effects of measurement heterogeneity[4,5]. The validation procedure was repeated 10,000 times for each simulation scenario. The number of events was around 1,000 in each dataset, which exceeds the minimal requirement for validation studies[26,31].

The simulation outcome measures were the average c-statistic, calibration slope, calibration-in-the-large coefficient, and Brier score. The c-statistic was computed using the somers2 function of the `rms` package[27]. The calibration slope was computed by regressing the observed outcome in the validation dataset on the linear predictor, as defined in Equation (9). We evaluated calibration graphically by plotting loess calibration curves and overlaying the plots of all 10 000 resamplings[26,32]. The calibration-in-the-large was computed as the intercept of the recalibration model, while using an offset for the linear predictor[1]. The empirical Brier score was computed using Equation (6). Additionally, we evaluated in-sample predictive performance as a reference for effects on out-of-sample performance.

5.2| Simulation results

Identical measurement error structures at derivation and validation resulted in consistent predictive performance across settings. All out-of-sample measures of predictive performance were affected by measurement heterogeneity. Effects on predictive performance measures were largest in the single-predictor model (Table 3). The two-predictor model in which one of the predictors was measured consistently over settings (Figure 3) outperformed the

model in which none of the predictors were measured consistently across settings (Figure 4). Inspection of calibration plots confirmed all patterns of miscalibration discussed below (Online Supplementary File 2). By and large, the impact of correlation between predictors on other parameters was minimal since the correlation structure was equal across compared settings, hence, we show combined results in the figures.

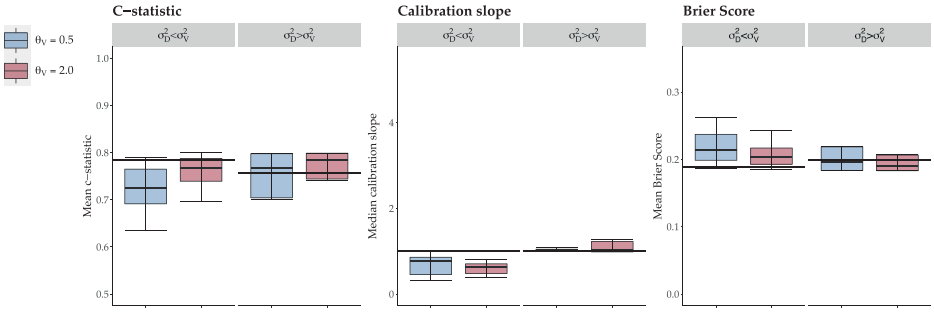


Figure 3. Measures of predictive performance under measurement heterogeneity in one of two predictors in finite sample simulations. Mean c-statistic, median calibration slope, and mean Brier score averaged over 10,000 repetitions with interquartile range and 95% confidence interval for a two-predictor model where one of the predictors is measured consistent across settings, whereas the other is measured heterogeneously. Horizontal bars indicate performance measures at model derivation, while boxes indicate performance at external validation. The predictor measurement structure in the derivation set (n = 2,000) corresponds to the random measurement error model (Equation (2)). In the validation set (n = 2,000), predictor measurements consist of varying structures under Equation (1).

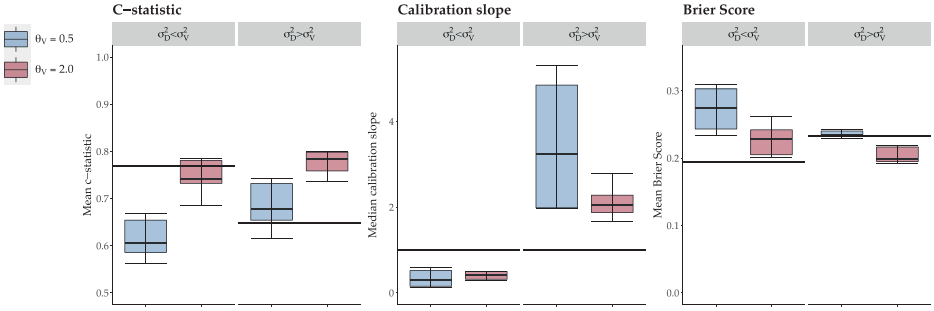


Figure 4. Measures of predictive performance under measurement heterogeneity in both predictors in finite sample simulations. Mean c-statistic, median calibration slope, and mean Brier score averaged over 10,000 repetitions with interquartile range and 95% confidence interval of a two-predictor logistic regression model in which both predictors are measured heterogeneously across settings. Horizontal bars indicate performance measures at model derivation, while boxes indicate performance at external validation. Measurements in the derivation set (n = 2,000) are recreated using Equation (2), which corresponds to the random measurement error model. In the validation set (n = 2,000), measurements correspond to various measurement

Table 3 Out-of-sample predictive performance measures under measurement heterogeneity in a single-predictor logistic regression model. Mean c-statistic, median calibration slope, mean calibration-in-the-large, and mean Brier score (standard deviation) at external validation of a single-predictor logistic regression model transported from a derivation set ($n = 2{,}000$) where measurement procedures were described by the random measurement error model (Equation (2)) to validation sets ($n = 2{,}000$) with various measurement structures under Equation (1). Predictive performance measures were averaged over 10,000 repetitions. All calibration slopes in the derivation set were equal to 1.0 (0.0) and are therefore not reported structures under Equation (1).

Measurement structure at validation		C-statistic		Calibration slope	Calibration-in-the-large (x 10)	Brier score	
		Derivation	Validation			Derivation	Validation
$\sigma^2_{\epsilon(D)} < \sigma^2_{\epsilon(V)}$	$\psi = 0, \theta = 0.5$	0.745 (0.033)	0.590 (0.034)	0.247 (0.153)	-0.002 (0.006)	0.204 (0.012)	0.281 (0.033)
	$\psi = 0, \theta = 1.0$	0.745 (0.033)	0.655 (0.045)	0.380 (0.180)	0.008 (0.014)	0.204 (0.012)	0.257 (0.031)
	$\psi = 0, \theta = 2.0$	0.745 (0.033)	0.726 (0.033)	0.428 (0.125)	-0.009 (0.003)	0.204 (0.012)	0.232 (0.023)
	$\psi = 0.25, \theta = 0.5$	0.745 (0.033)	0.589 (0.034)	0.247 (0.153)	-2.202 (0.643)	0.204 (0.012)	0.283 (0.032)
	$\psi = 0.25, \theta = 1.0$	0.745 (0.033)	0.655 (0.045)	0.380 (0.180)	-2.210 (0.652)	0.204 (0.012)	0.258 (0.031)
	$\psi = 0.25, \theta = 2.0$	0.745 (0.033)	0.726 (0.033)	0.428 (0.125)	-2.205 (0.651)	0.204 (0.012)	0.233 (0.023)
$\sigma^2_{\epsilon(D)} = \sigma^2_{\epsilon(V)}$	$\psi = 0, \theta = 0.5$	0.700 (0.068)	0.635 (0.069)	0.812 (0.291)	0.001 (0.006)	0.217 (0.020)	0.235 (0.015)
	$\psi = 0, \theta = 1.0$	0.700 (0.068)	0.700 (0.068)	1.000 (0.000)	0.001 (0.008)	0.217 (0.020)	0.218 (0.020)
	$\psi = 0, \theta = 2.0$	0.700 (0.068)	0.753 (0.042)	0.955 (0.377)	-0.002 (0.013)	0.217 (0.020)	0.204 (0.014)
	$\psi = 0.25, \theta = 0.5$	0.700 (0.068)	0.635 (0.069)	0.811 (0.293)	-1.529 (1.027)	0.217 (0.020)	0.237 (0.014)
	$\psi = 0.25, \theta = 1.0$	0.700 (0.068)	0.700 (0.068)	1.002 (0.002)	-1.530 (1.033)	0.217 (0.020)	0.219 (0.019)
	$\psi = 0.25, \theta = 2.0$	0.700 (0.068)	0.753 (0.042)	0.955 (0.377)	-1.526 (1.024)	0.217 (0.020)	0.205 (0.013)
$\sigma^2_{\epsilon(D)} > \sigma^2_{\epsilon(V)}$	$\psi = 0, \theta = 0.5$	0.655 (0.045)	0.681 (0.045)	3.147 (1.991)	0.003 (0.007)	0.230 (0.011)	0.234 (0.009)
	$\psi = 0, \theta = 1.0$	0.655 (0.045)	0.745 (0.034)	3.106 (1.563)	0.000 (0.006)	0.230 (0.011)	0.220 (0.014)
	$\psi = 0, \theta = 2.0$	0.655 (0.045)	0.781 (0.014)	2.160 (0.969)	0.005 (0.009)	0.230 (0.011)	0.203 (0.013)
	$\psi = 0.25, \theta = 0.5$	0.655 (0.045)	0.681 (0.045)	3.156 (2.001)	-0.846 (0.528)	0.230 (0.011)	0.235 (0.008)
	$\psi = 0.25, \theta = 1.0$	0.655 (0.045)	0.745 (0.034)	3.102 (1.559)	-0.846 (0.532)	0.230 (0.011)	0.221 (0.013)
	$\psi = 0.25, \theta = 2.0$	0.655 (0.045)	0.781 (0.014)	2.159 (0.967)	-0.851 (0.535)	0.230 (0.011)	0.203 (0.013)

5.2.1 Random measurement heterogeneity

When measurements were less precise at validation compared to derivation, i.e., when $\sigma^2_{\epsilon(D)} < \sigma^2_{\epsilon(V)}$, the c-statistic decreased, and Brier score increased at validation. In the single-predictor model, the c-statistic decreased from 0.75 at derivation to 0.59 – 0.73 at validation and the Brier score increased from 0.20 at derivation to 0.23 – 0.28 at validation (Table 3, bottom rows). Furthermore, the median calibration slope at validation was smaller than 1, ranging from 0.25 – 0.43 in the single-predictor model. When measurements were more precise at validation compared to derivation, i.e., when $\sigma^2_{\epsilon(D)} > \sigma^2_{\epsilon(V)}$, the c-statistic was increased, from 0.66 to 0.68 – 0.78 in the single-predictor model, and the Brier score was decreased, changing from 0.23 to 0.20 – 0.24 in the single-predictor model. However, the improved c-statistic and Brier score were accompanied by median calibration slopes greater than 1, ranging from 2.16 – 3.16 in the single-predictor model (Table 3, top rows). Calibration-in-the-large was not affected by random measurement heterogeneity. Similar effects on predictive performance were observed for the two-predictor models, which are presented graphically in Figures 3 and 4.

5.2.2 Systematic measurement heterogeneity

When measurements at external validation changed by a constant compared to derivation, i.e., when $\psi_D = 0$ and $\psi_V = 0.25$, the risk on observing the outcome was systematically overestimated, which is reflected in the negative value for calibration-in-the-large coefficient (Table 3). Changes in ψ had little effect on the calibration slope and Brier score, and no apparent effect on the c-statistic. Multiplicative systematic measurement heterogeneity, i.e., $\theta_D \neq \theta_V$, reinforced or counterbalanced effects of random measurement heterogeneity in the direction of the systematic measurement heterogeneity. When the association between x and w was relatively weak at validation, e.g., when $\theta_V = 0.5$, predictive performance deteriorated (blue bars in Figures 3 and 4), whereas predictive performance improved when the association between x and w was relatively strong, e.g., when $\theta_V = 2.0$ (red bars in Figures 3 and 4).

5.2.3 Differential measurement heterogeneity

We highlight four specific scenarios in which the single-predictor model was derived under differential random measurement error, i.e., $\sigma^2_{\epsilon 1} \neq \sigma^2_{\epsilon_0}$, and validated using nondifferential measurements, and vice versa (Table 4). Differential measurement led to miscalibration at external validation in all scenarios. The c-statistic and Brier score at validation slightly improved when cases were measured less precise at derivation

or more precise at validation. For example, when cases were measured less precise at derivation, i.e., $\sigma^2_{\epsilon_1(D)} > \sigma^2_{\epsilon_0(D)}$, the c-statistic increased from 0.66 to 0.71 at validation and the Brier score decreased from 0.23 to 0.22. However, the median calibration slope at validation was 1.86.

Table 4 Effects of differential measurement of predictors in events and non-events in four scenarios. Mean c-statistic, median calibration slope, and mean Brier score (standard deviation) averaged over 10 000 repetitions for a single-predictor logistic regression model under four specific measurement error structures varying in the degree of random measurement variance under the differential measurement error model (Equation 1). By default, σ^2_{ϵ} is set to 1.0. When $\sigma^2_{\epsilon 1} = 0.5$, measurements are more precise in cases. When $\sigma^2_{\epsilon 1} = 2.0$, measurements are less precise in cases

Differential measurement error at…		C-statistic		Calibration slope	Brier score	
		Derivation	Validation		Derivation	Validation
Derivation	$\sigma^2_{\epsilon 1} = 0.5$	0.730 (0.011)	0.707 (0.012)	0.780 (0.071)	0.209 (0.004)	0.219 (0.004)
	$\sigma^2_{\epsilon 1} = 2.0$	0.655 (0.012)	0.707 (0.012)	1.856 (0.208)	0.231 (0.003)	0.223 (0.002)
Validation	$\sigma^2_{\epsilon 1} = 0.5$	0.706 (0.012)	0.730 (0.011)	1.293 (0.120)	0.217 (0.003)	0.211 (0.003)
	$\sigma^2_{\epsilon 1} = 2.0$	0.706 (0.012)	0.655 (0.012)	0.547 (0.061)	0.217 (0.004)	0.237 (0.005)

6 | Discussion

Heterogeneity of predictor measurements across settings can have a substantial impact on the out-of-sample performance of a prediction model. When predictor measurements are more precise at derivation compared to validation, model discrimination and accuracy at validation deteriorate, and the provided predicted probabilities are too extreme, similar to when a model is overfitted with respect to the derivation data. When predictor measurements are less precise at derivation compared to validation, discrimination and accuracy at validation tend to improve, but the provided predicted probabilities are too close to the outcome prevalence, similar to statistical underfitting. These key findings of our study are summarized in Table 5. The current study emphasizes that a prediction model not only concerns the algorithm relating predictors to the outcome, but also depends on the procedures by which model input is measured, i.e., qualitative differences in data collection.

Measurement error is commonly thought not to affect the validity of prediction models, based on the general idea that unbiased associations between predictor and outcome are no prerequisite in prediction studies[18]. By taking the measurement error perspective, our study revealed that prediction research requires consideration

of variation in measurement procedures *across* different settings of derivation and validation, rather than analyzing the amount of measurement error *within* a study. A recent systematic review by Whittle and colleagues demonstrated that measurement error was not acknowledged in many prediction studies and pointed out the need to investigate consequences of measurement error in prediction research[33]. An important starting point for this research following from our study is that the generalizability of prediction models depends on the transportability of measurement structures.

Table 5 Key Findings. Effects of measurement heterogeneity on predictive performance in general scenarios of measurement heterogeneity.

Predictor measurements at validation		**Predictive performance at validation**			
		Discrimination	**Calibration-in-the-large**	**Calibration slope**	**Overall accuracy**
Less precise compared to derivation;	$\sigma^2_{\epsilon(D)} < \sigma^2_{\epsilon(V)}$	Deteriorated	-	$b<1$	Deteriorated
More precise compared to derivation;	$\sigma^2_{\epsilon(D)} > \sigma^2_{\epsilon(V)}$	Improved	-	$b>1$	Improved
Weaker association with actual predictor value, while					
- less precise compared to derivation;	$\theta_V > 1.0$, $\sigma^2_{\epsilon(D)} < \sigma^2_{\epsilon(V)}$	Stronger deterioration	-	Stronger $b<1$	Stronger deterioration
- more precise compared to derivation;	$\theta_V > 1.0$, $\sigma^2_{\epsilon(D)} > \sigma^2_{\epsilon(V)}$	Less improvement	-	Stronger $b>1$	Less improvement
Stronger association with actual predictor value, while					
- less precise compared to derivation;	$\theta_V > 1.0$, $\sigma^2_{\epsilon(D)} < \sigma^2_{\epsilon(V)}$	Less deterioration	-	Less $b<1$	Less deterioration
- more precise compared to derivation;	$\theta_V > 1.0$, $\sigma^2_{\epsilon(D)} > \sigma^2_{\epsilon(V)}$	Stronger improvement	-	Less $b>1$	Stronger improvement
Increased by a constant relative to derivation.	$\psi_V > 0$	-	$a<0$	-	-

Specification of measurement heterogeneity can help explaining discrepancies in predictive performance between derivation and validation setting in a pragmatic way. The relatedness between derivation and validation samples is generally quantified in terms of similarity in person-characteristics (also referred to as "case-mix"), and regression coefficients[1]. Previously proposed measures to express sample relatedness are the mean and spread of the linear predictor[7] or the correlation structure of predictors in both samples[34]. The information on sample relatedness can be incorporated in benchmark values of predictive performance to assess model transportability[6]. While regression coefficients and case-mix distributions clearly quantify sample relatedness, it is impossible to disentangle the sources of discrepancies from these statistical measures. For example, a decrease of the regression coefficients or the spread of the linear predictor at external validation could be due to differences across settings in either person-characteristics or the means by which these characteristics were measured. Moreover, less precise predictor measurements affect both the regression coefficients and the spread of the linear predictor, meaning that measurement heterogeneity can mask similarities and differences between the individuals in a derivation and validation sample. Knowledge of substantive differences between derivation and validation setting can help researchers determining to which extent the prediction model is transportable.

In theory, measurement error correction procedures could be applied to adjust for measurement heterogeneity when data on both X and W are available[16]. Alternatively, the degree of measurement heterogeneity could be quantified using the residual intraclass correlation (RICC), which expresses the clustering of measurements across physicians or centers[9]. Yet, we expect that the applicability of these methods in correcting for measurement heterogeneity will be limited not only due to the fact that individual patient data of both the derivation and validation set are required, but furthermore because it is infeasible to disentangle measurement parameters from other characteristics of the data. The main contribution of the taxonomy of measurement error models rises from its aptitude to conceptualize measurement heterogeneity across settings in pragmatic terms.

The following implications for prediction studies follow from our work. Ideally, prediction models are derived from predictor measurements that resemble measurement procedures in the intended setting of application. Data collection protocols that reduce measurement error to a minimum do not necessarily benefit the performance of the model as the precision of measurements will most likely not be

obtained in validation (or application) settings. Deriving a prediction model from these precise measurements could result in miscalibration similar to model overfitting and reduced discrimination and accuracy at external validation. Furthermore, researchers should bear in mind the implications of using a "readily available dataset" for model derivation or validation as data quality directly affects predictive performance of the model. For instance, validating a model in a clinical trial dataset, in which measurements typically contain minimal measurement error, may increase measures of discrimination and accuracy, yet the model may provide predicted probabilities too close to the event rate due to miscalibration. Another example is the promising use of large routine care datasets for model validation[5,35,36]. Predictor measurement procedures may vary greatly within such datasets or differ from the procedures used to collect the data for the derivation study, which could increase the predictor measurement variance to a level that no longer resembles the amount of measurement variance within a clinical setting. Hence, rather than analyzing data because they are available, prediction models should be derived from and validated on datasets collected with measurement procedures that are in widespread use in the intended clinical setting. Finally, it is important to clearly report which measurement procedures were used for derivation or validation of a prediction model. The influential TRIPOD Statement has drawn attention to the importance of reporting measurement procedures[8]. Our findings indicate that descriptions of measurement procedures at model derivation are essential for proper external validation of the model. Likewise, the validation studies ideally contain descriptions of deviations from measurements used at derivation, as these may introduce discrepancies in predictive performance.

Our study redefines the importance of predictor measurements in the context of prediction research. We highlight heterogeneity in predictor measurement procedures across settings as an important driver of unanticipated predictive performance at external validation. Preventing measurement heterogeneity at the design phase of a prediction study, both in development and validation studies, facilitates interpretation of predictive performance, and benefits the transportability of the prediction model.

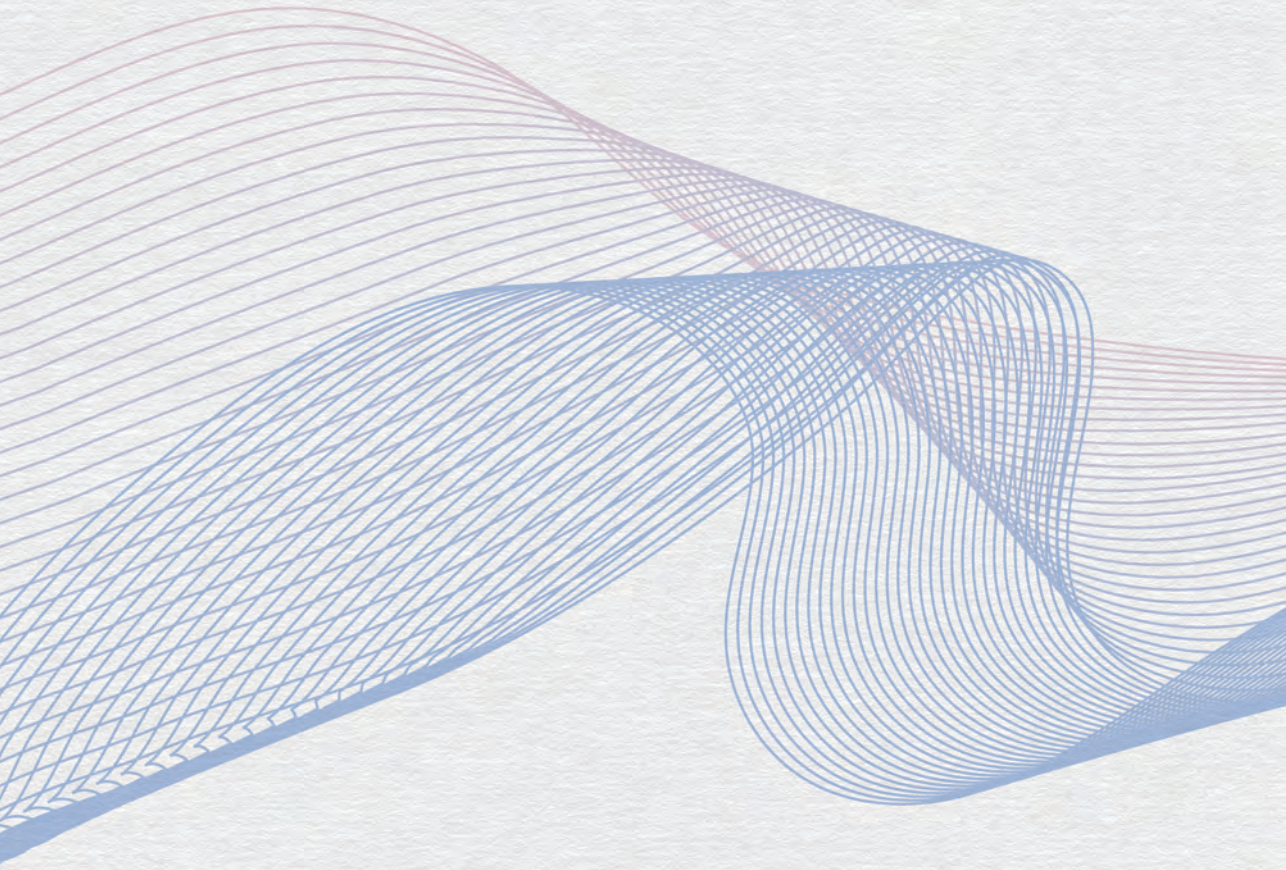
Online Supplementary Files

The supplementary files referred to in this Chapter are available online at https://doi.org/10.1002/sim.8183

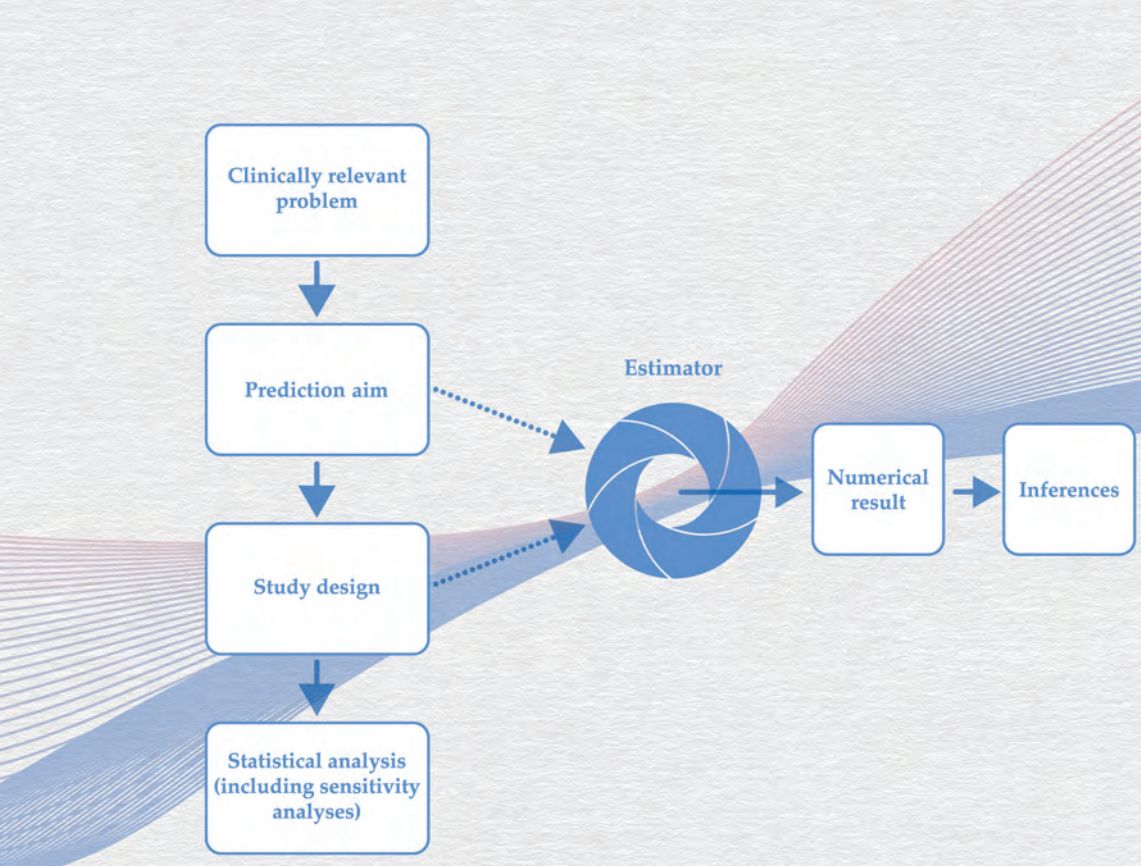
References

1. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Springer Science & Business Media; 2008.
2. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine*. 2000;19(4):453-473.
3. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in Medicine.* 2016;35(2):214-226.
4. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine.* 2004;23(16):2567-2586.
5. Steyerberg EW, Uno H, Ioannidis JP, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. *Journal of Clinical Epidemiology.* 2018;98:133-143.
6. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology.* 2010;172(8):971-980.
7. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine.* 2013;32(18):3158-3180.
8. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *British Medical Journal*. 2015;350:g7594.
9. Wynants L, Timmerman D, Bourne T, Van Huffel S, Van Calster B. Screening for data clustering in multicenter studies: the residual intraclass correlation. *BMC Medical Research Methodology.* 2013;13(1):128.
10. Mikula A, Hetzel S, Binkley N, Anderson P. Clinical height measurements are unreliable: a call for improvement. *Osteoporosis International.* 2016;27(10):3041-3047.
11. Drawz PE, Ix JH. BP measurement in clinical practice: time to SPRINT to guideline-recommended protocols. *Journal of the American Society of Nephrology.* 2018;29(2):383-388.
12. Genders TS, Steyerberg EW, Hunink MM, et al. Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts. *British Medical Journal*. 2012;344.
13. Aubert CE, Folly A, Mancinetti M, Hayoz D, Donzé J. Prospective validation and adaptation of the HOSPITAL score to predict high risk of unplanned readmission of medical patients. *Swiss Medical Weekly.* 2016;146:w14335.
14. Herder GJ, Van Tinteren H, Golding RP, et al. Clinical prediction model to characterize pulmonary nodules: validation and added value of 18 F-fluorodeoxyglucose positron emission tomography. *Chest.* 2005;128(4):2490-2496.
15. Al-Ameri A, Malhotra P, Thygesen H, et al. Risk of malignancy in pulmonary nodules: a validation study of four prediction models. *Lung Cancer.* 2015;89(1):27-30.
16. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statistics in Medicine.* 2014;33(12):2137-2155.
17. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass).* 2010;21(1):128.
18. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement error in nonlinear models: a modern perspective.* Chapman and Hall/CRC; 2006.
19. White E. Measurement error in biomarkers: sources, assessment, and impact on studies. *IARC Scientific Publications.* 2011(163):143-161.

20. Sackett D. Bias in analytic research In: The Case-Control Study Consensus and Controversy. *Journals of Chronic Diseases*. 1979:51-63.
21. Khudyakov P, Gorfine M, Zucker D, Spiegelman D. The impact of covariate measurement error on risk prediction. *Statistics in Medicine*. 2015;34(15):2353-2367.
22. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Medical Research Methodology*. 2012;12(1):1-8.
23. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. 1950;78(1):1-3.
24. Blattenberger G, Lad F. Separating the Brier score into calibration and refinement components: A graphical exposition. *The American Statistician*. 1985;39(1):26-32.
25. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*. 1986;5(5):421-433.
26. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*. 2016;74:167-176.
27. Harrell Jr FE. *rms: regression modeling strategies*. R package version 3.6-3. 2013.
28. Cox DR. Two further applications of a model for binary regression. *Biometrika*. 1958;45(3/4):562-565.
29. Rosella LC, Corey P, Stukel TA, Mustard C, Hux J, Manuel DG. The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model. *Population Health Metrics*. 2012;10(1):1-11.
30. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020.
31. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology*. 2005;58(5):475-483.
32. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*. 2014;33(3):517-535.
33. Whittle R, Royle K-L, Jordan KP, Riley RD, Mallen CD, Peat G. Prognosis research ideally should measure time-varying predictors at their intended moment of use. *Diagnostic and Prognostic Research*. 2017;1(1):1-9.
34. Kundu S, Mazumdar M, Ferket B. Impact of correlation of predictors on discrimination of risk models in development and external populations. *BMC Medical Research Methodology*. 2017;17(1):1-9.
35. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *British Medical Journal*. 2016;353.
36. Cook JA, Collins GS. The rise of big clinical databases. *Journal of British Surgery*. 2015;102(2):e93-e101.



7


Clinically relevant problem
Prediction aim
Study design
Statistical analysis (including sensitivity analyses)
Estimator
Numerical result
Inferences

Changing predictor measurement procedures affected the performance of prediction models in clinical examples

The objective of this study was to quantify the impact of predictor measurement heterogeneity on prediction model performance. Predictor measurement heterogeneity refers to variation in the measurement of predictor(s) between the derivation of a prediction model and its validation or application. It arises, for instance, when predictors are measured using different measurement instruments or protocols. We examined effects of various scenarios of predictor measurement heterogeneity in real-world clinical examples using previously developed prediction models for diagnosis of ovarian cancer, mutation carriers for Lynch syndrome, and intrauterine pregnancy. Changing the measurement procedure of a predictor influenced the performance at validation of the prediction models in nine clinical examples. Notably, it induced model miscalibration. The calibration intercept at validation ranged from -0.70 to 1.43 (0 for good calibration), while the calibration slope ranged from 0.50 to 1.67 (1 for good calibration). The difference in c-statistic and scaled Brier score between derivation and validation ranged from -0.08 to +0.08 and from -0.40 to +0.16, respectively. This study illustrates that predictor measurement heterogeneity can influence the performance of a prediction model substantially, underlining that predictor measurements used in research settings should resemble clinical practice. Specification of measurement heterogeneity can help researchers explaining discrepancies in predictive performance between derivation and validation setting.

This chapter was based on: Luijken K, Wynants L, van Smeden M, Van Calster B, Steyerberg EW, Groenwold RHH. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *Journal of Clinical Epidemiology*. 2020;119:7-18.

1 | Background

Clinical prediction models are commonly applied in clinical practice to assist healthcare professionals in determining a patient's diagnosis or prognosis[1]. Clinical prediction models are applied to patients that were not part of the data used to derive the model, often with the aim to estimate a probability for the presence of a disease or future health state[2]. When applied on new patients, the performance in estimating these probabilities is often different from the performance in the derivation data. This is commonly explained by model overfitting with respect to the derivation data[3-6] and differences in patient characteristics (case-mix) between derivation and validation settings[7-9].

Previous studies have identified imprecise predictor measurement procedures as another reason for a suboptimal performance of prediction models at derivation[10,11] and highlighted that differences in predictor measurement procedures between derivation and validation setting substantially affected performance at validation[12-14]. Predictor variables may be measured by different procedures in external validation data than those applied in derivation data, that is, according to different measurement protocols, measurement instruments or by applying different predictor definitions. We refer to these differences in measurement across settings as predictor measurement heterogeneity. Simulation studies have shown that predictor measurement heterogeneity can induce miscalibration of prediction models and affect discrimination and accuracy at external validation[12]. While predictor measurement heterogeneity across derivation and validation samples appears to be common in clinical (research) settings (see for example [4,15,16]), its impact on the performance of prediction models at validation is not well-studied using empirical data.

In this study, we quantify the impact of predictor measurement heterogeneity on predictive performance in a series of real-world clinical examples.

2 | Illustrating and defining predictor measurement heterogeneity

We briefly illustrate predictor measurement heterogeneity here using measurements of the predictor body mass index (BMI). We fitted a logistic regression model to predict the presence of pre-stage diabetes containing only two parameters for a linear and a quadratic term of BMI besides the intercept (this example was adapted from Rosella and colleagues[11]). Data were available on 1,264 participants from the NHANES Study 2013-

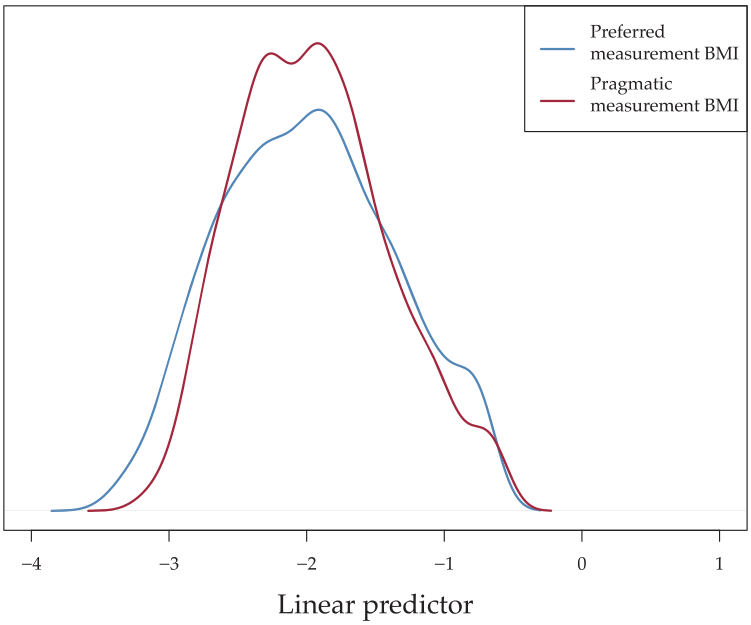


Figure 1. Impact of predictor measurement heterogeneity on distributions of linear predictors. Density of the logit transformation of the predicted risks (linear predictor) from a logistic regression model predicting the probability of a pre-stage of diabetes using the predictor BMI. BMI was obtained as an instrumental (preferred) and self-reported (pragmatic) measure. Distributions of the linear predictors for both procedures are presented. The prediction model was: $\text{logit}\left(P(Y_i = 1|BMI_i)\right) = \beta_0 + \beta_1 BMI_i + \beta_2 BMI_i{}^2$

2014[17]. BMI data were computed from participants' height and weight measurements, obtained by a trained examiner who followed a standardized protocol[18]. Since this measurement is close to what we would consider the ideal method of measurement, we will refer to it as the *preferred measurement*. A second measurement of BMI was computed via self-reported weight and height by the participants, which we will refer to as the *pragmatic measurement*. The concept *predictor measurement heterogeneity* refers to the phenomenon where the predictor measurement strategy at derivation differs from the measurement strategy at validation or application of the prediction model.

A second regression model was fitted with a linear and quadratic term for BMI using the pragmatic measurement of BMI. Comparing the output of the two regression models, it becomes clear that substituting the preferred measurement of BMI with the pragmatic measurement changed the distribution of the linear predictor (Figure 1). To better understand how substitution of pragmatic by preferred measurements (and vice versa) can affect predictive performance, we present empirical case studies in the next sections.

3 | Methods

We examined effects of predictor measurement heterogeneity in previously established prediction models, using three empirical datasets on the diagnosis of ovarian cancer, hereditary non-polyposis colorectal cancer (Lynch syndrome), and intrauterine pregnancy, respectively. Scenarios from various clinical domains were investigated in order to provide a general assessment of the potential impact of predictor measurement heterogeneity.

3.1 | Example dataset 1: diagnosis of ovarian cancer

The International Ovarian Tumor Analysis (IOTA) dataset includes clinical and ultrasound information on 5,914 non-pregnant women with at least one persistent adnexal mass[19]. We used data from IOTA phases I-III (1999-2012) in which we studied two prediction models, here referred to as Model 1 and Model 2. Model 1 is a logistic regression model that estimates the probability of presence of ovarian mass malignancy from pre-operatively measured predictors: age (years); maximal diameter of the tumor (mm); personal history of ovarian cancer (yes/no); current use of hormonal therapy (yes/no); experience of pain during examination (yes/no); presence of ascites (yes/no); presence of blood flow within a solid papillary projection (yes/no); maximal diameter of the largest solid component (mm); presence of irregular cyst walls (yes/no); presence of acoustic shadows (yes/no); color score of intratumoral blood flow (ordinal, ranging 1-4); and presence of entirely solid tumors (yes/no). Model 1 is based on the LR1 model, which was developed and internally validated in IOTA phase-I data[20] and has been externally validated several times[21-23]. Model 2 is a logistic regression model to preoperatively diagnose ovarian mass malignancy by age (years), the proportion of solid tissue, the presence of more than ten locules (yes/no), the number of papillary structures, the presence of acoustic shadows (yes/no), and the presence of ascites (yes/no). It is a previously described reduction of Model 1, developed for methodological illustrations[24].

3.2 | Example dataset 2: prediction of mutation carrier status (Lynch syndrome)

We analysed data from 19,866 patients with colorectal cancer (CRC), who were tested for mutations in Lynch syndrome–related mismatch repair genes. We studied a simplification of the $PREMM_{1,2}$-model[25] and MMRpredict model[5,26] in the Lynch syndrome dataset, which we refer to as Model 3. Model 3 is a logistic regression model that predicts the prevalence of MLH1/MSH2 mutations from the following predictors measured at baseline: sex; age at CRC diagnosis (years); and family history of CRC

and endometrial cancer. Family history was defined as a weighted sum of positive first- and second-degree relatives, where second-degree relatives were weighted half times the first-degree relatives. The sum ranged from 0-3, with family history coded as 0, 1, or 2+ affected relatives.

3.3 | Example dataset 3: prediction of intrauterine pregnancy

We analysed data from 75 consecutive patients at the Early Pregnancy and Acute Gynaecology Unit (EPAGU) at Queen Charlottes' and Chelsea Hospital from November 2013 to May 2014. We studied a logistic regression model in the pregnancy data, here referred to as Model 4, that predicts the probability of an ongoing intrauterine pregnancy based on measurements of hCG level at presentation (pmol/L) and an hCG ratio of hCG at 48h after presentation to hCG at presentation. hCG Levels could be measured using two different measurement instruments, named the 'ria kit' and the 'imm kit'. Model 4 is adapted from an existing multinomial logistic regression model (named M4)[27], by grouping the outcome categories 'ectopic pregnancy' and 'pregnancy of unknown location'.

3.4 | Models and assessment of predictive performance

To separate the impact of predictor measurement heterogeneity from other possible external validation effects on predictive performance, such as changes in case-mix and outcome incidence, we focus on derivation and validation within the same study population and evaluate predictive performance[28]. In each example, we defined scenarios of measurement heterogeneity by identifying two measurement procedures of a single predictor: a *preferred* measurement and a *pragmatic* measurement (Table 1). The terms '*preferred*' and '*pragmatic*' are only meant in a relative sense: a preferred measurement may still be far from the ideal measurement of a particular phenomenon, but as a predictor of a particular outcome it could be preferable over the pragmatic measurement in terms of a lower measurement error or anticipated better predictive potential for the particular outcome.

For each scenario, we assessed the optimism-corrected predictive performance of a regular maximum likelihood logistic regression model under both predictor measurement *homogeneity* and *heterogeneity*. The optimism correction was performed since measures of predictive performance based on the derivation data may give an over-optimistic assessment of model performance, as maximum likelihood models are generated to provide the best fit for the derivation data[28]. Measures of predictive performance were obtained by deriving and validating a prediction model in 500 bootstrap samples and averaging optimism-corrected measures of performance over the bootstrap samples

Table 1 Scenarios of measurement heterogeneity in four clinical prediction models

Scenario	Dataset	Model	Measurement heterogeneity		Explanation
			Preferred procedure (scale)	Pragmatic procedure (scale)	
1	IOTA	1	Maximal diameter tumor (continuous)	Mean diameter tumor (continuous)	In the original model, the diameter of the tumor was measured as the maximum of three measurements of the tumor lesion in different dimensions. Alternatively, the mean of these three measurements could be used as model input.
2	IOTA	1	Maximal diameter solid component tumor, non-truncated (continuous)	Mean diameter solid component tumor, non-truncated (continuous)	In the original model, the diameter of the largest solid component of the tumor was measured as the maximum of three measurements of the solid component of the tumor lesion in different dimensions. Alternatively, the mean of these three measurements could be used as model input.
3	IOTA	1	Diameter solid component truncated at 50mm (continuous)	Original diameter solid component (continuous)	The diameter of the largest solid component of the tumor was truncated at 50 mm in the original model. In application of this model, the truncation can be ignored or forgotten.
4	IOTA	1	Color-score (ordinal, 1-4)	Color-score at extremes (dichotomous, 1 or 4)	The intratumoral blood flow was scored by a colorscore ranging 1-4 in the original model. Alternatively, the extremes of this score (1 or 4) could be used as model input, because: - A colorscore is a subjective measurement; at model application, physicians could score the colors at the extremes (either no or high blood flow). - Researchers could use a (public) dataset for model validation in which only a binary version of the score is available, rather than a categorical score, and recode this variable into scores 1 or 4.
5	IOTA	2	≥10 locules (binary)	≥5 locules (binary)	The original model included a dichotomized version of the number of locules, where the cut-off was at 10 locules. At model validation or application, the cut-off value for dichotomization could be different.

Table 1 *(continued)*

Scenario	Dataset	Model	Measurement heterogeneity		Explanation
			Preferred procedure (scale)	Pragmatic procedure (scale)	
6	Lynch syndrome	3	Family history CRC summarized by counting 0,1,2+ FDRs and 0,1,2+ SDR, weighted by a half (categorical, 0-3)	Family history CRC summarized by counting only FDRs (categorical, 0-3)	Family history of CRC is computed as a weighted count of diagnoses of CRC in first- and second-degree relatives. Possibly, the history of CRC is recorded for first-degree relatives only and used as model input.
7	Lynch syndrome	3	Family history EC summarized by counting 0,1,2+ FDRs and 0,1,2+ SDR, weighted by a half (categorical, 0-3)	Family history EC summarized by counting only FDRs (categorical, 0-3)	Family history of EC is computed as a weighted count of diagnoses of EC in first- and second-degree relatives. Possibly, the history of EC is recorded for first-degree relatives only and used as model input.
8	Pregnancy	4	hCG level measured in serum, using the ria kit (continuous)	hCG level measured in urine, using the ria kit (continuous)	A hCG measurement is preferably obtained from serum samples but could alternatively be obtained from urine samples.
9	Pregnancy	4	hCG level measured in serum, using the ria kit (continuous)	hCG level measured in serum, using the imm kit (continuous)	A hCG measurement can be obtained using different measurement kits, e.g., the ria kit or imm kit.

Abbreviations: CRC = colorectal cancer, EC = endometrial cancer, FDR = first-degree relative, SDR = second-degree relative.

(see Online Supplementary File 1 for detailed explanation)[28]. To assess predictor measurement homogeneity, the prediction model was derived and validated based on the same predictor definitions. To assess predictor measurement heterogeneity, a derivation and validation setting were recreated by deriving the model using the *preferred* measurement and validating the model using the *pragmatic* measurement, denoted scenario 1a-9a, or by deriving the model using the *pragmatic* measurement and validating the model using the *preferred* measurement, denoted scenario 1b-9b. Note that we isolate the impact of measurement heterogeneity here by keeping all other factors besides measurement heterogeneity constant (i.e., the modelling strategy, included predictors, and patient characteristics are equal at derivation and validation).

Measures of predictive performance were the calibration-in-the-large coefficient and calibration slope from a logistic recalibration model, the c-statistic (area under the ROC curve) and the Brier score. Model calibration refers to the agreement between observed outcomes and risk estimates[1,29]. The calibration-in-the-large coefficient evaluates whether there is a difference between the observed event fraction and the average predicted risk (0 for perfect calibration) and is estimated as the intercept of the recalibration model while the calibration slope is fixed at a value of 1. The calibration slope (<1 indicating overfitting, i.e., predicted risks that are too extreme, and >1 indicating underfitting) was computed by regressing the observed outcome on the logit transformation of the predicted risks and evaluated graphically by plotting lowess calibration curves. We considered the scaled Brier score, in which the Brier score is scaled by its maximum score under a non-informative model, $Brier_{scaled} = 1 - Brier / Brier_{max}$, so that it ranges from 0 for perfect predictions to 1 for non-informative predictions[1,29].

To quantify the resemblance between the predictor measurement procedures, the partial correlation between the *preferred* and *pragmatic* predictors was estimated by correlating residuals of two linear regression models regressing each of the predictor measurements on the outcome and other covariates in the model. Shrunken regression coefficients from a Ridge logistic regression model were estimated, for which the tuning parameter (necessary for shrinkage) was determined by the value minimizing the deviance in 10-fold cross-validation[30]. All analyses were performed in R 3.5.1.[31] and R code is available at https://github.com/KLuijken/Prediction_Measurement_Heterogeneity_Examples. Measures of predictive performance were obtained using the `rms` package[32].

4 | Results

Measures of predictive performance in all scenarios are presented in Table 2 (under measurement homogeneity) and Table 3 (under measurement heterogeneity). The latter results are presented graphically in Figure 2-4, for the IOTA dataset, Lynch syndrome dataset and pregnancy dataset, respectively. Each scenario is discussed in detail in the Online Supplementary File 2. Results after regression shrinkage (Ridge regression, Online Supplementary File 3) did not differ from results without; we only discuss the latter here.

4.1 | Predictive performance under measurement homogeneity

Measures of predictive performance varied between models. However, within scenarios a switch in measurement strategy for a single predictor did not materially impact the predictive performance (Table 2), with the exception of scenario 8, where the c-statistic and scaled Brier score decreased when the *pragmatic* measurement was used (pregnancy dataset, $N = 75$).

4.2 | Predictive performance under measurement heterogeneity

Table 3 shows estimates of predictive performance under predictor measurement heterogeneity across the different models. The calibration-in-the-large coefficient at validation ranged from -0.70 (95% CI, -1.26; -0.21) to 1.43 (95% CI, 0.31; 2.54), suggesting systematic over- or underestimation of the predicted risks. The calibration slope at validation ranged from 0.50 (95% CI, 0.22; 0.91) to 1.67 (95% CI, 0.83; 3.47), consistent with overfitting (too extreme predictions) and underfitting (too narrow range of predictions), respectively. The differences in c-statistic between derivation and validation were small to moderate, ranging from -0.08 (95% CI, -0.11; -0.07) to +0.08 (95% CI, 0.07; 0.11). The change in scaled Brier score between derivation and validation ranged from -0.40 (95% CI, -0.80; -0.15) to +0.16 (95% CI, 0.15; 0.17). In what follows, we provide a detailed discussion of predictive performance, where we group the scenarios by type of predictor measurement heterogeneity.

In settings where a different measure of aggregation for defining the predictor was used (scenario 1ab-2ab), the direction of miscalibration was related to the shift in aggregational measure (Figure 2). When the maximum tumor diameter was used at derivation and the mean at validation, the calibration-in-the-large coefficient was larger than zero, indicating systematic underestimation of the predicted risks at validation

(scenario 1a, 2a). The reverse occurred in scenario 1b and 2b. Calibration-in-the-large was more strongly affected in scenario 2ab, where the predictor-outcome association was higher than in scenario 1ab.

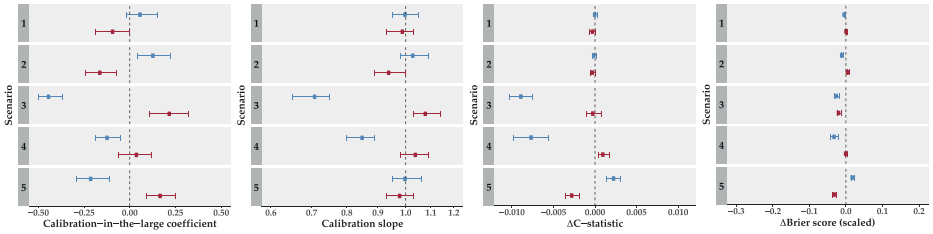


Figure 2. Measures of predictive performance under predictor measurement heterogeneity of a model predicting the probability of having ovarian mass malignancy. The model is applied to the International Ovarian Tumor Analysis (IOTA) dataset, containing information on 5,914 non-pregnant women (1999-2012). Error bars represent the 95-percentile interval over 500 bootstrap samples. Blue error bars indicate scenario 1a-5a, meaning the model was derived using the preferred measurement and validated using the pragmatic measurement, red error bars indicate scenario 1b-5b, meaning the model was derived using the pragmatic measurement and validated using the preferred measurement.

Truncation of a continuous predictor measurement showed the following effects on calibration (scenario 3ab; Figure 2). When the truncated value was used for model derivation and the non-truncated value at validation, the calibration-in-the-large coefficient indicated systematic overestimation of the predicted risks at validation, and the calibration slope was smaller than one, indicating overfitting with respect to the derivation data; predicted risks were too extreme compared to the observed proportions (and vice versa in scenario 3b).

When the categories of an ordinal predictor were collapsed into a binary variable by using only the extremes of the scale (scenario 4a; Figure 2), the calibration-in-the-large coefficient indicated systematic overestimation of the predicted risks, the calibration slope indicated overfitting with respect to the derivation data, and the c-statistic decreased (and vice versa in scenario 4b).

When a more stringent dichotomization was used at validation by shifting the cut-off of a count upward (scenario 5b; Figure 2) or including only first-degree relatives in a summary score on family history, rather than both first- and second-degree relatives (scenario 6a,7a; Figure 3), risks were systematically underestimated, as indicated by the calibration-in-the-large coefficient (and vice versa in 5a, 6b, 7b). In scenario 6a, the calibration slope indicated model underfitting, the c-statistic decreased, and the scaled Brier score decreased (and vice versa in scenario 6b).

Switching from serum to urine hCG measurements (scenario 8ab) showed the following effects on predictive performance (Figure 4). When the predictor measurement had a smaller variance at derivation compared to validation (scenario 8a), the calibration-in-the-large coefficient indicated systematic overestimation of the predicted risks and the calibration slope indicated model overfitting. The c-statistic and scaled Brier score decreased. The reverse occurred when the predictor measurement had lower variance at validation compared to derivation (scenario 8b), except for the scaled Brier score, which decreased again.

A switch in measurement instrument, i.e., using the ria kit versus using the imm kit for hCG measurement in serum (scenario 9ab; Figure 4), minimally affected predictive performance. The large uncertainty around measures of predictive performance in scenario 8ab and 9ab can largely be explained by the limited sample size.

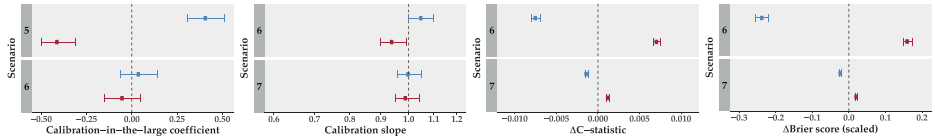


Figure 3. Measures of predictive performance under predictor measurement heterogeneity of a model predicting the probability of having Lynch syndrome–related mismatch repair genes. The model is applied to the Lynch syndrome dataset, containing information on 19,866 patients with colorectal cancer who were tested for mutations. Error bars represent the 95-percentile interval over 500 bootstrap samples. Blue error bars indicate scenario 6a and 7a, meaning the model was derived using the preferred measurement and validated using the pragmatic measurement, red error bars indicate scenario 6b and 7b, meaning the model was derived using the pragmatic measurement and validated using the preferred measurement.

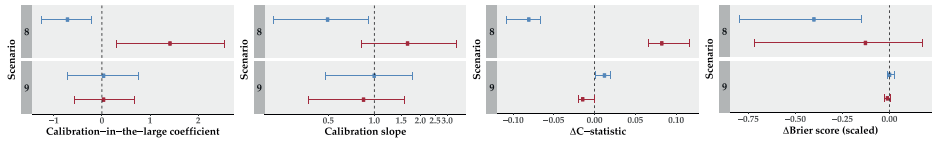


Figure 4. Measures of predictive performance under predictor measurement heterogeneity of a model predicting the probability of intrauterine pregnancy. The model is applied to the pregnancy dataset, containing information on 75 patients at the Early Pregnancy and Acute Gynaecology Unit (EPAGU) at Queen Charlottes' and Chelsea Hospital (2013 – 2014). Error bars represent the 95-percentile interval over 500 bootstrap samples. Blue error bars indicate scenario 8a and 9a, meaning the model was derived using the preferred measurement and validated using the pragmatic measurement, red error bars indicate scenario 8b and 9b, meaning the model was derived using the pragmatic measurement and validated using the preferred measurement.

Table 2 Measures of optimism-corrected predictive performance under predictor measurement homogeneity

Scenario	Event fraction	Measurement strategy		Mean value measurement	Standard deviation measurement	C-statistic	Scaled Brier score
1	0.33	Preferred	Maximal diameter tumor	82.06	52.07	0.94 (0.94; 0.95)	0.62 (0.60; 0.64)
	0.33	Pragmatic	Mean diameter tumor	68.98	42.68	0.94 (0.94; 0.95)	0.62 (0.60; 0.65)
2	0.33	Preferred	Maximal diameter solid component tumor, non-truncated	27.55	39.34	0.94 (0.93; 0.95)	0.60 (0.58; 0.62)
	0.33	Pragmatic	Mean diameter solid component tumor, non-truncated	22.73	32.71	0.94 (0.93; 0.95)	0.60 (0.58; 0.62)
3	0.33	Preferred	Diameter solid component truncated at 50mm	18.91	21.31	0.94 (0.94; 0.95)	0.62 (0.60; 0.64)
	0.33	Pragmatic	Original diameter solid component	27.55	39.34	0.94 (0.93; 0.95)	0.60 (0.58; 0.62)
4	0.33	Preferred	Color-score 4 categories	2.25	0.99	0.94 (0.94; 0.95)	0.62 (0.60; 0.64)
	0.33	Pragmatic	Color-score dichotomous	2.20	1.47	0.94 (0.94; 0.95)	0.61 (0.59; 0.64)
5	0.33	Preferred	≥10 locules	0.08	0.27	0.89 (0.89; 0.90)	0.46 (0.44; 0.49)
	0.33	Pragmatic	≥5 locules	0.19	0.40	0.90 (0.89; 0.91)	0.47 (0.44; 0.49)
6	0.10	Preferred	Family history CRC both FDR and SDR	0.64	0.76	0.78 (0.77; 0.79)	0.16 (0.14; 0.17)
	0.10	Pragmatic	Family history CRC FDR only	0.45	0.66	0.77 (0.76; 0.78)	0.14 (0.13; 0.16)
7	0.10	Preferred	Family history EC both FDR and SDR	0.10	0.31	0.78 (0.77; 0.79)	0.16 (0.14; 0.17)
	0.10	Pragmatic	Family history EC FDR only	0.07	0.28	0.78 (0.77; 0.79)	0.16 (0.14; 0.17)

Table 2 *(continued)*

Scenario	Event fraction	Measurement strategy		Mean value measurement	Standard deviation measurement	C-statistic	Scaled Brier score
8*	0.40	Preferred	hCG level measured in serum, ria kit	2.74 and -0.23	1.44 and 0.86	0.90 (0.81; 0.97)	0.54 (0.32; 0.78)
	0.40	Pragmatic	hCG level measured in urine, ria kit	4.32 and -0.16	1.74 and 1.12	0.81 (0.70; 0.91)	0.27 (0.05; 0.52)
9*	0.40	Preferred	hCG level measured in serum, ria kit	2.74 and -0.23	1.44 and 0.86	0.90 (0.81; 0.97)	0.54 (0.31; 0.78)
	0.40	Pragmatic	hCG level measured in serum, imm kit	2.90 and -0.27	1.41 and 0.84	0.91 (0.83; 0.98)	0.56 (0.32; 0.79)

Measures of predictive performance were averaged over 500 bootstrap samples and corrected for optimism. Confidence intervals for the c-statistic and scaled Brier score were obtained by subtracting the optimism from the 95-percentile interval over the 500 bootstrap estimates of the performance measure under predictor measurement homogeneity. Scaled Brier score is computed as: $1 - Brier/Brier_{max}$.

Abbreviations: CRC = colorectal cancer, EC = endometrial cancer, FDR = first-degree relative, SDR = second-degree relative.

* The hCG measurements are included in the model as a log-transformed hCG measurement at presentation plus a log-transformed ratio of hCG at 48h to hCG-at-presentation measurement.

Table 3 Measures of predictive performance under predictor measurement heterogeneity

Scenario	Measurement strategy at derivation	Measurement strategy at validation	ρ_{part}	Calibration-in-the-large	Calibration slope	ΔC-statistic * 100	Δ scaled Brier score
1a	Maximal diameter tumor	Mean diameter tumor	0.98	0.06 (-0.02; 0.15)	1.00 (0.95; 1.05)	0.00 (-0.00; 0.02)	-0.00 (-0.00; -0.00)
1b	Mean diameter tumor	Maximal diameter tumor		-0.09 (-0.19; -0.00)	0.99 (0.93; 1.03)	-0.02 (-0.06; 0.00)	0.00 (0.00; 0.00)
2a	Maximal diameter solid component tumor, non-truncated	Mean diameter solid component tumor, non-truncated	0.98	0.13 (0.04; 0.22)	1.03 (0.98; 1.09)	-0.00 (-0.00; 0.00)	-0.01 (-0.01; -0.01)
2b	Mean diameter solid component tumor, non-truncated	Maximal diameter solid component tumor, non-truncated		-0.16 (-0.24; -0.07)	0.94 (0.89; 1.00)	-0.00 (-0.00; 0.00)	0.01 (0.00; 0.01)
3a	Diameter solid component truncated at 50mm	Original diameter solid component	0.65	-0.44 (-0.50; -0.37)	0.71 (0.65; 0.75)	-0.88 (-1.02; -0.75)	-0.02 (-0.03; -0.02)
3b	Original diameter solid component	Diameter solid component truncated at 50mm		0.22 (0.11; 0.32)	1.08 (1.03; 1.14)	-0.02 (-0.10; 0.07)	-0.02 (-0.02; -0.01)
4a	Color-score 4 categories	Color-score dichotomous	0.81	-0.12 (-0.19; -0.05)	0.85 (0.80; 0.89)	-0.76 (-0.98; -0.56)	-0.03 (-0.04; -0.02)
4b	Color-score dichotomous	Color-score 4 categories		0.04 (-0.06; 0.12)	1.04 (0.98; 1.09)	0.10 (0.04; 0.17)	0.00 (-0.00; 0.00)
5a	≥10 locules	≥5 locules	0.56	-0.21 (-0.29; -0.11)	1.00 (0.95; 1.06)	0.23 (0.14; 0.31)	0.02 (0.01; 0.02)
5b	≥5 locules	≥10 locules		0.17 (0.09; 0.25)	0.98 (0.93; 1.03)	-0.27 (-0.35; - 0.18)	-0.03 (-0.04; -0.03)
6a	Family history CRC both FDR and SDR	Family history CRC FDR only	0.90	0.41 (0.31; 0.51)	1.05 (1.00; 1.10)	-0.74 (-0.79; -0.70)	-0.24 (-0.26; -0.22)
6b	Family history CRC FDR only	Family history CRC both FDR and SDR		-0.41 (-0.50; -0.31)	0.94 (0.90; 0.99)	0.71 (0.67; 0.76)	0.16 (0.15; 0.17)
7a	Family history EC both FDR and SDR	Family history EC FDR only	0.93	0.04 (-0.06; 0.14)	1.00 (0.96; 1.05)	-0.13 (-0.15; -0.12)	-0.02 (-0.02; -0.02)

Table 3 *(continued)*

Scenario	Measurement strategy at derivation	Measurement strategy at validation	ρ_{part}	Calibration-in-the-large	Calibration slope	ΔC-statistic * 100	Δ scaled Brier score
7b	Family history EC FDR only	Family history EC both FDR and SDR		-0.05 (-0.15;0.05)	0.99 (0.95; 1.04)	0.12 (0.11; 0.14)	0.02 (0.02; 0.03)
8a*	hCG level measured in serum, using the ria kit	hCG level measured in urine, using the ria kit	0.61 and 0.92	-0.70 (-1.26; -0.21)	0.50 (0.22; 0.91)	-7.97 (-10.80; -6.59)	-0.40 (-0.80; -0.15)
8b*	hCG level measured in urine, using the ria kit	hCG level measured in serum, using the ria kit		1.43 (0.31; 2.54)	1.67 (0.83; 3.47)	8.34 (6.67; 11.63)	-0.12 (-0.72; 0.17)
9a*	hCG level measured in serum, using the ria kit	hCG level measured in serum, using the imm kit	0.98 and 0.997	0.05 (-0.71; 0.75)	1.01 (0.48; 1.78)	1.31 (0.07; 2.00)	0.00 (-0.01; 0.03)
9b*	hCG level measured in serum, using the imm kit	hCG level measured in serum, using the ria kit		0.05 (-0.56; 0.68)	0.86 (0.37; 1.58)	-1.36 (-2.02; 0.00)	-0.01(-0.03; 0.01)

Performance measures under predictor measurement heterogeneity: median calibration coefficients and mean difference scores of c-statistic and scaled Brier score over 500 bootstrap samples with 95-percentile intervals. Δ indicates that the measure of predictive performance under predictor measurement homogeneity is subtracted from the performance measure under predictor measurement heterogeneity. Scaled Brier score is computed as: $1 - Brier/Brier_{max}$.
Abbreviations: CRC = colorectal cancer, EC = endometrial cancer, FDR = first-degree relative, SDR = second-degree relative.
* The hCG measurements are included in the model as a log-transformed hCG measurement at presentation plus a log-transformed ratio of hCG at 48h to hCG-at-presentation measurement.

5 | Discussion

In this study, we evaluated the impact of predictor measurement heterogeneity in nine different scenarios in three clinical datasets. A change in measurement strategy of a predictor within the derivation set, from *preferred* measurement to *pragmatic* measurement or vice versa, minimally affected measures of predictive performance in our example studies. We found that heterogeneity of measurements across settings of derivation and validation can have a substantial impact on the performance of a prediction model, most notably on overall accuracy and calibration of risk predictions, resulting in systematic over- or under estimation of predicted risks and risk models that are consistent with overfitting (systematically too extreme predictions) or underfitting (systematically a too narrow range of predictions).

In the examples, the impact on calibration was larger when predictors were stronger associated with the outcome or when the partial correlation between predictor measurement strategies was lower. Using Ridge regression as a shrinkage method or correcting for optimism did not compensate for effects of measurement heterogeneity in our study. The variety of effects on predictive performance in the examples illustrated the difficulty of anticipating the exact impact of predictor measurement heterogeneity, emphasising the need to be generally mindful of (dis)similarities of predictor measurement strategies between derivation and validation studies.

We observed small effects of predictor measurement heterogeneity on the discriminatory power of the model at validation in our examples. Previous simulation studies found larger effects on the c-statistic[10-12]. Our finding may be explained by the fact that we focused on within-sample predictive performance under measurement heterogeneity in a single predictor. With a larger number of predictors subject to measurement heterogeneity, we anticipate the combined effect on the discrimination performance can be larger. Also, given that the c-statistic is a rank order statistic it is possible that this metric is less affected by measurement heterogeneity[33].

Our findings showed that internal predictive performance may not be affected by changes in predictor measurement strategy within the same dataset, in line with previous studies[10,11,34]. Previous research showed that variations in measurement error did not affect risk calibration[10], but these findings were restricted to within-sample effects on predictive performance only. Within the derivation dataset, models derived using logistic regression achieve, by definition, a calibration-in-the-large coefficient of zero and calibration slope of one regardless of the measurement error structure of

predictors[29]. Our study highlights that this does not apply when the degree or structure of measurement error varies across settings of derivation and validation, the case of measurement heterogeneity.

It is common practice in validation studies to quantify the relatedness of derivation and validation samples by inspecting the distribution of the linear predictors, also referred to as comparison of *case-mix* distributions[8,9,35]. Dissimilarities in the distributions of the linear predictor between derivation and validation may rise from both actual differences in patient characteristics and differences in the procedures used to measure patient characteristics. By identifying predictor measurement heterogeneity as a separate explanation of discrepancies in linear-predictor distributions across settings, our findings can facilitate implementation of the influential TRIPOD statement in clinical prediction research[36].

Our study has several limitations. Firstly, it was limited to three empirical datasets with a diagnostic outcome modelled using logistic regression. One dataset, from the IOTA study, was a multicenter study in which homogeneous measurement strategies across centers was among its hallmark characteristics, see[19]. Measurement heterogeneity within development and validation studies, e.g., due to variability in measurement precision between clinicians or centres[37], is an important topic for future research. Given the potential impact and limited attention to date[38], research is needed on the effect of measurement heterogeneity for other statistical models and outcomes (e.g., survival models for time-to-event outcomes) and the impact on more flexible prediction modelling strategies. Finally, the similarity between the preferred and pragmatic measurement of a predictor was quantified using a partial correlation coefficient. This measure quantifies the conditional association between predictor measurements rather than agreement[39]. Since the current paper aimed to examine whether variation in predictor measurement strategies across settings can have an effect on predictive performance of any degree or direction, we presented a single measurement of similarity of predictor measurements and left out further quantification. One way to visualize agreement between measurement could be Bland-Altman plots[40].

The following recommendations follow from our work. When a prediction model is derived, predictor measurements should be clearly defined and ideally resemble procedures in the intended setting of application as closely as possible. For prediction model validation studies, we encourage researchers to investigate to which extent predictor measurement procedures are homogeneous and may have contributed to differences in predictive performance between the validation and derivation setting.

Accurate reporting of predictor measurement heterogeneity in both derivation and validation studies is therefore essential. Furthermore, we take the position that addressing measurement heterogeneity at the data collection stage is preferred over statistical correction for measurement error in predictors. Corrections – typically aiming to alleviate measurement-error bias in regression coefficients – may increase rather than reduce the measurement heterogeneity[12].

We emphasize that consideration of predictor measurement heterogeneity is crucial also in the implementation stage of a prediction model in clinical practice. Deployment of a prediction model might alter predictor measurement heterogeneity. For example, after implementation of a prediction model, physicians may be recommended to use a more precise or standardized measurement (or even routinely measure predictors that were not measured in all patients up to that point). For implementation of prediction models in clinical practice, our findings indicate that measurement procedures should follow the measurements in derivation and validation datasets as closely as possible.

In summary, our findings highlight that predictor measurement heterogeneity can have a substantial influence on the performance of a prediction model, most notably on risk calibration. Explicit reporting of the procedures and timing involved in measurement of predictors in derivation and validation studies is vital to improve the performance and applicability of prediction models in clinical practice.

Online Supplementary Files

The supplementary files referred to in this Chapter are available online at https://doi.org/10.1016/j.jclinepi.2019.11.001

References

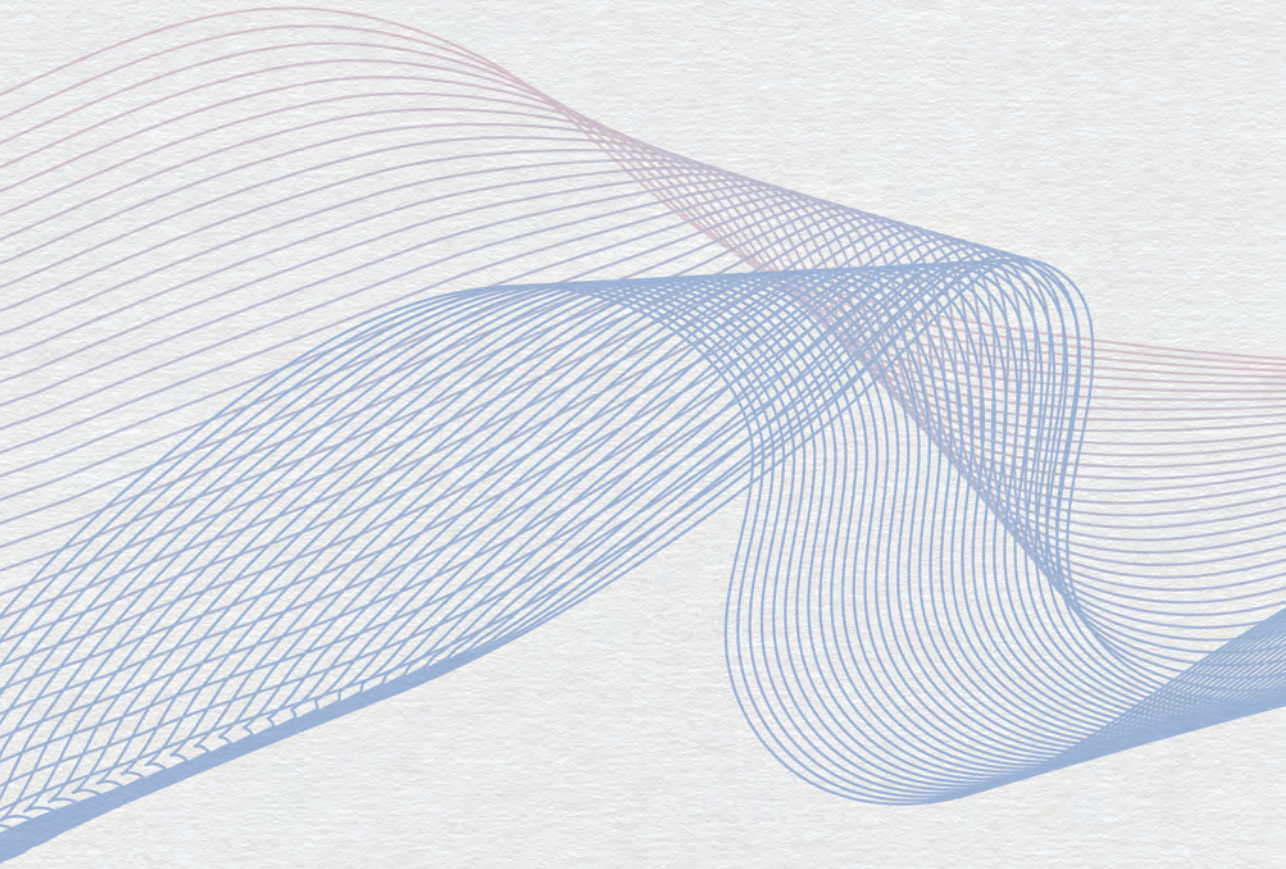
1. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Springer Science & Business Media; 2008.
2. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine.* 2000;19(4):453-473.
3. Wynants L, Collins GS, Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG: An International Journal of Obstetrics & Gynaecology.* 2017;124(3):423-432.
4. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *British Medical Journal.* 2015;350:g7594.
5. Steyerberg EW, Uno H, Ioannidis JP, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. *Journal of Clinical Epidemiology.* 2018;98:133-143.
6. Toll D, Janssen K, Vergouwe Y, Moons K. Validation, updating and impact of clinical prediction rules: a review. *Journal of Clinical Epidemiology.* 2008;61(11):1085-1094.
7. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine.* 1999;130(6):515-524.
8. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine.* 2013;32(18):3158-3180.
9. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology.* 2010;172(8):971-980.
10. Khudyakov P, Gorfine M, Zucker D, Spiegelman D. The impact of covariate measurement error on risk prediction. *Statistics in Medicine.* 2015;34(15):2353-2367.
11. Rosella LC, Corey P, Stukel TA, Mustard C, Hux J, Manuel DG. The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model. *Population Health Metrics.* 2012;10(1):20.
12. Luijken K, Groenwold RH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. *Statistics in Medicine.* 2019;38(18):3444-3459.
13. Pajouheshnia R, Van Smeden M, Peelen L, Groenwold R. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *Journal of Clinical Epidemiology.* 2019;105:136-141.
14. Pajouheshnia R, Groenwold RH, Peelen LM, Reitsma JB, Moons KG. When and how to use data from randomised trials to develop or validate prognostic models. *British Medical Journal.* 2019;365:l2154.
15. Te Velde E, Nieboer D, Lintsen A, et al. Comparison of two models predicting IVF success; the effect of time trends on model performance. *Human Reproduction.* 2013;29(1):57-64.
16. Smith T, Muller DC, Moons KG, et al. Comparison of prognostic models to predict the occurrence of colorectal cancer in asymptomatic individuals: a systematic literature review and external validation in the EPIC and UK Biobank prospective cohort studies. *Gut.* 2019;68(4):672-683.
17. Control CfD, Prevention. National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Questionnaire (or Examination Protocol, or Laboratory Protocol). *http://wwwcdcgov/nchs/nhaneshtm.* 2006.
18. Control CfD, Prevention. National health and nutrition examination survey (NHANES): Anthropometry procedures manual. *National Center for Health Statistics, editor Atlanta, GA: Centers for Disease Control.* 2007.

19. Timmerman D, Valentin L, Bourne T, Collins W, Verrelst H, Vergote I. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*. 2000;16(5):500-505.
20. Timmerman D, Testa AC, Bourne T, et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. *Journal of Clinical Oncology*. 2005;23(34):8794-8801.
21. Van Holsbeke C, Van Calster B, Testa AC, et al. Prospective internal validation of mathematical models to predict malignancy in adnexal masses: results from the international ovarian tumor analysis study. *Clinical Cancer Research*. 2009;15(2):684-691.
22. Timmerman D, Van Calster B, Testa AC, et al. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. *Ultrasound in Obstetrics and Gynecology*. 2010;36(2):226-234.
23. Nunes N, Ambler G, Hoo W-L, et al. A prospective validation of the IOTA logistic regression models (LR1 and LR2) in comparison to subjective pattern recognition for the diagnosis of ovarian cancer. *International Journal of Gynecological Cancer*. 2013;23(9):1583-1589.
24. Wynants L, Vergouwe Y, Van Huffel S, Timmerman D, Van Calster B. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. *Statistical Methods in Medical Research*. 2018;27(6):1723-1736.
25. Balmaña J, Stockwell DH, Steyerberg EW, et al. Prediction of MLH1 and MSH2 mutations in Lynch syndrome. *Journal of the American Medical Association*. 2006;296(12):1469-1478.
26. Barnetson RA, Tenesa A, Farrington SM, et al. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. *New England Journal of Medicine*. 2006;354(26):2751-2763.
27. Van Calster B, Abdallah Y, Guha S, et al. Rationalizing the management of pregnancies of unknown location: temporal and external validation of a risk prediction model on 1962 pregnancies. *Human Reproduction*. 2013;28(3):609-616.
28. Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans M, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*. 2001;54(8):774-781.
29. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*. 2016;74:167-176.
30. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 1992;41(1):191-201.
31. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020.
32. Harrell Jr FE. *rms: regression modeling strategies*. R package version 5.1-4. 2016.
33. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clinical Chemistry*. 2008;54(1):17-23.
34. Carroll RJ, Ruppert D, Crainiceanu CM, Stefanski LA. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC; 2006.
35. Kundu S, Mazumdar M, Ferket B. Impact of correlation of predictors on discrimination of risk models in development and external populations. *BMC Medical Research Methodology*. 2017;17(1):63.
36. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine*. 2015;162(1):W1-W73.
37. Wynants L, Timmerman D, Bourne T, Van Huffel S, Van Calster B. Screening for data clustering in multicenter studies: the residual intraclass correlation. *BMC Medical Research Methodology*. 2013;13(1):128.

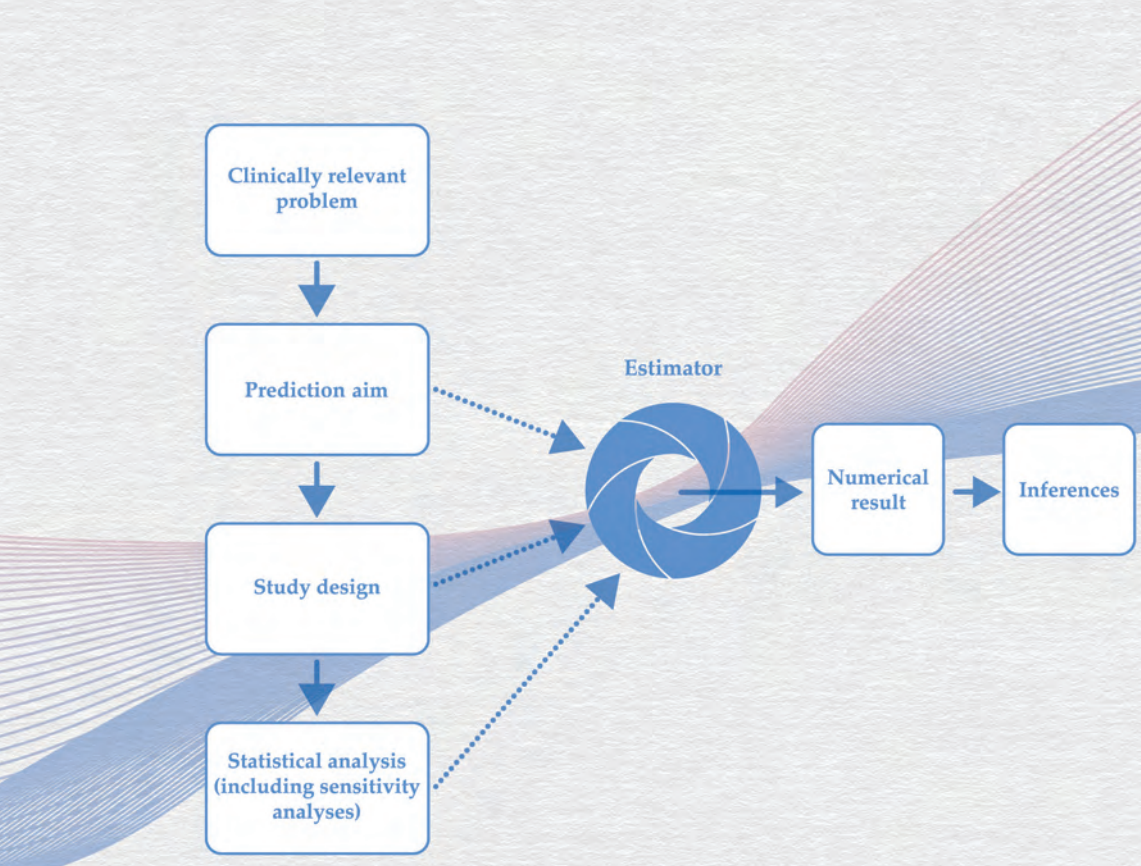
38. Whittle R, Peat G, Belcher J, Collins GS, Riley RD. Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported. *Journal of Clinical Epidemiology*. 2018;102:38-49.

39. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*. 1996;1(1):30.

40. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*. 1986;327(8476):307-310.



8


Clinically relevant problem
Prediction aim
Study design
Statistical analysis (including sensitivity analyses)
Estimator
Numerical result
Inferences

Quantitative prediction analysis to investigate predictive performance under predictor measurement heterogeneity at model implementation

When a predictor variable is measured in similar ways at the derivation and validation setting of a prognostic prediction model, yet both differ from the intended use of the model in practice (i.e., 'predictor measurement heterogeneity'), performance of the model at implementation needs to be inferred. This study proposed an analysis to quantify the impact of anticipated predictor measurement heterogeneity. A simulation study was conducted to assess the impact of predictor measurement heterogeneity across validation and implementation setting in time-to-event outcome data. The use of the quantitative prediction analysis was illustrated using an example of predicting the risk of developing type-2 diabetes with heterogeneity in measurement of the predictor body mass index. In the simulation study, calibration-in-the-large of prediction models was poor and overall accuracy was reduced in all scenarios of predictor measurement heterogeneity. Model discrimination decreased with increasing random predictor measurement heterogeneity. Heterogeneity of predictor measurements across settings of validation and implementation reduced predictive performance at implementation of prognostic models with a time-to-event outcome. When validating a prognostic model, the targeted clinical setting needs to be considered and analyses can be conducted to quantify the anticipated impact of predictor measurement heterogeneity on model performance at implementation.

This chapter was based on: Luijken K, Song J, Groenwold RHH, Quantitative prediction error analysis to investigate predictive performance under predictor measurement heterogeneity at model implementation. *Diagnostic and Prognostic Research* (in press).

1 | Background

Clinical prognostic models aim to provide predictions of an outcome for individuals who have not been part of the modelling process[1-5]. The quantity that a clinical prediction model targets is defined by specifying the outcome, (candidate) predictors, population, setting, time of prediction, and prediction horizon as specifically as possible[6]. When the research setting does not correspond to the intended setting of application in clinical practice[7,8] or when modelling strategies are inappropriate[9,10], the predictive performance of a prognostic model may be suboptimal at implementation.

One reason for suboptimal predictive performance of a model at implementation are differences in predictor measurement procedures between model development and implementation in practice[7,11]. When discrepancies in predictor measurement procedures impact the performance of a clinical prediction model, this is referred to as *predictor measurement heterogeneity*[12]. The impact of predictor measurement heterogeneity on predictive performance at external validation has been quantified for models of binary outcome data[11-14] and illustrated in empirical data sets for logistic regression diagnostic prediction models[11,15]. However, the step towards model implementation in a target population has not been studied yet. The impact of predictor measurement heterogeneity in time-to-event data has not received adequate attention either.

In the current study, we suggest an approach to anticipate the impact of predictor measurement heterogeneity on a prognostic model when it is implemented in clinical practice. We assess the impact of predictor measurement heterogeneity in time-to-event outcome data using large-sample simulations. We propose a quantitative prediction analysis for validation studies that can be used to quantify the impact of anticipated predictor measurement heterogeneity in one of the predictors. This is illustrated using an example of a model predicting the 6-year risk of developing type-2 diabetes.

2 | Predictor measurement heterogeneity

For a prognostic model to provide correct predictions of an outcome in a clinical setting, several phases of model development should be considered, which is outlined in Figure 1[5,16-18]. Ideally, a prognostic model is derived using data that corresponds to the targeted implementation setting (derivation setting)[19,20]. Predictive performance is typically evaluated by measures of apparent performance and measures of performance after internal validation of the model, i.e., after correcting for optimism about the

performance[21,22]. When the internal predictive performance of the model is sufficient, its performance can be investigated using external data (validation setting)[23,24], which is preferably done multiple times[25-27]. When predictive performance at external validation is sufficiently well, implementation of the model in clinical practice could be considered (implementation setting), advisably after performing an impact analysis[28,29].

One aspect to consider in all phases of development of a prognostic model is predictor measurement heterogeneity, indicated in the blue box in Figure 1. Procedures to collect and measure predictor data for derivation and validation studies ideally correspond to the future implementation setting. When predictor measurement procedures at derivation and/or validation deviate from the predictor measurement procedure used in clinical practice, this can affect the predictive performance at implementation.

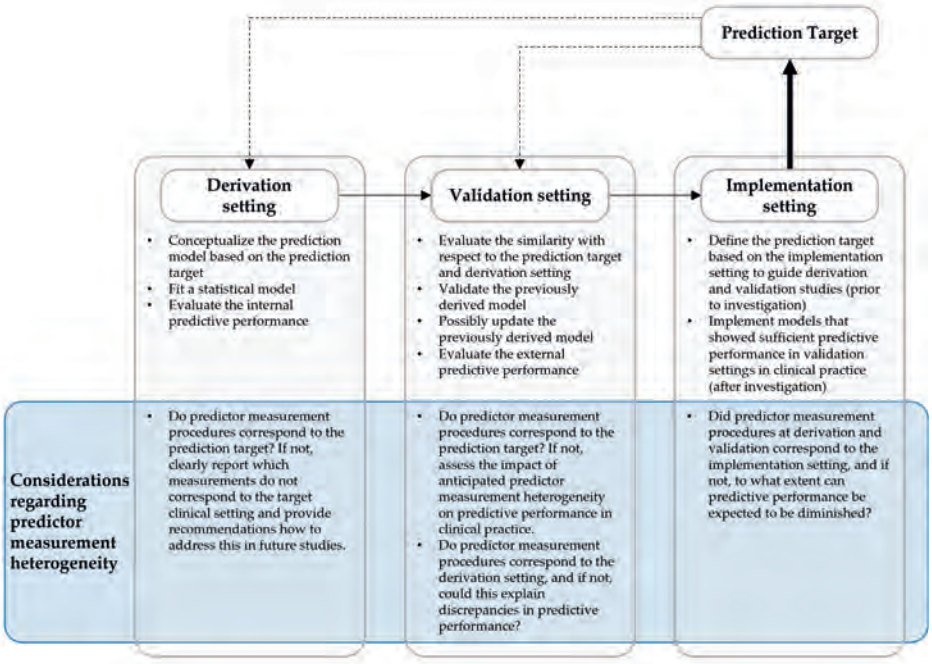


Figure 1. An overview of the derivation, validation, and implementation setting of a prognostic model, highlighting considerations regarding predictor measurement heterogeneity. Note that 'impact analysis' research is a phase between validation and implementation that is not addressed in this diagram. A prediction target is defined by specifying the target population, setting, outcome, (candidate) predictors, time of prediction, and prediction horizon as specifically as possible.

3 | Simulation study

We performed a simulation study to investigate the impact of predictor measurement heterogeneity across validation and implementation setting on out-of-sample predictive performance of a survival model derived and validated in time-to-event outcome data. We assumed that all other possible sources of discrepancy in predictive performance are not present, e.g., there are no differences in outcome prevalence and treatment assignment policy, there is no overfitting with respect to the derivation data, and the prognostic model is correctly specified in terms of functional form and included interactions. We used (very) large samples (n = 1,000,000) to minimize the role of random simulation error.

3.1 | Design of simulation study

Online Supplement 1 contains a detailed description of the simulation study. The main aspects of the design of the simulations study are described below and are reported following previous recommendations[30].

Data-generating mechanism: We simulated derivation, validation, and implementation data sets with 1,000,000 observations containing a continuous predictor variable X from a standard normal distribution. A time-to-event outcome was simulated for each subject so that outcomes followed a Cox-exponential model, using methods described by Bender and colleagues[31] (see Table 1 for simulation parameters). We generated data sets without censoring (median survival time $t = 6.6$). Additionally, data sets with administrative censoring after $t = 15$ (74% event fraction, median survival time 6.6) and with random censoring (69% event fraction, median survival time $t = 5.6$) were generated.

At implementation, a different measurement of predictor X was available, denoted W. Predictor measurement heterogeneity across validation and implementation setting was recreated using measurement error models, similar to[12]. The mean difference between X and W was denoted ψ (additive systematic measurement heterogeneity), the linear association between X and W was denoted θ (multiplicative systematic measurement heterogeneity), and the variance introduced by random deviations from X was denoted σ^2_ϵ, where non-zero values of σ^2_ϵ reflect that measurement W is less precise than X (random measurement heterogeneity).

In total, 162 scenarios were evaluated (27 scenarios of predictor measurement heterogeneity, for 2 different models under 3 different censoring mechanisms).

Prediction target: The prediction target was defined as obtaining correct predictions of the outcome risk at time point $t = 6.5$ conditional on predictor measurement W measured at moment of prediction (i.e., at $t = 0$).

Table 1 Simulation parameters.

Parameter	Value
Baseline hazard of an event	0.1
Conditional hazard ratio for association predictor *X* and survival times	2
Time point of administrative censoring	15
Baseline hazard of censoring	0.01
Conditional hazard ratio for association between random variable for censoring and censoring times	3
Mean of predictor *X* and random variable for censoring	0
Variance of predictor *X* and random variable for censoring	1
Predictor *W* at implementation*	
ψ	-0.3 to 0.3
θ	0.5 to 2
σ_ϵ	0 to √2

* At implementation, a different measurement of predictor *X* was available, denoted measurement *W*. The connection between *X* and *W* was defined using the following measurement heterogeneity model: $\mathbb{E}(W) = \psi + \theta\mathbb{E}(X) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and where ψ denotes an additive shift in *W* with respect to *X*, θ denotes a multiplicative linear association between *W* and *X*, and σ_ϵ^2 denotes random deviations from *X*.

Methods: A parametric exponential survival model and a semi-parametric Cox regression model were fitted in the derivation data set. Although a prognostic model is typically internally validated before performing external validation[1,21], we did not perform an internal validation since issues of overfitting were expected to be negligible due to the large sample sizes. The prognostic model was externally validated at time $t = 6.5$ (around median survival time) under predictor measurement homogeneity in an independent (validation) data set. Predictor measurement homogeneity refers to the situation in which predictors are measured in the same way at derivation and validation. Furthermore, the predictive performance of the prognostic model was investigated in various implementation settings under predictor measurement heterogeneity. The procedure was performed once in each scenario.

Performance metrics: Predictive performance was evaluated at $t = 6.5$, i.e., approximately at the median survival time. Calibration of the model on average,

or 'calibration in the large'[32,33], was evaluated by the ratio of the observed marginal survival at $t = 6.5$ (obtained through a Kaplan-Meier curve) versus the predicted marginal survival at $t = 6.5$ (obtained by averaging predicted survival at $t = 6.5$ of each observation), denoted the observed / expected ratio (O/E ratio). Discrimination was evaluated by the cumulative-dynamic time-dependent area under the receiver operating characteristic curve AUC(t)[34-36]. Overall accuracy was evaluated by the index of prediction accuracy at $t = 6.5$, IPA(t), which equals a Brier score[37] at $t = 6.5$ that is benchmarked to a null model ignoring all patient specific information and simply predicts the empirical prevalence to each patient[38]. A perfect model has an IPA of 1, a non-informative model has an IPA of 0 and a negative IPA indicates a harmful model.

Software: The simulation study was performed using R statistical software version 3.6.3[39]. The simulation code is available from https://github.com/KLuijken/PMH_Survival.

3.2 | Results of simulation study

Predictor measurement heterogeneity affected predictive performance at implementation. In all scenarios of predictor measurement heterogeneity, the prognostic models were miscalibrated in the large (range O/E ratio 0.89 to 1.19, compared to 1.00 under predictor measurement homogeneity), and overall accuracy was reduced (range IPA(6.5) -0.17 to 0.17, compared to 0.17 under predictor measurement homogeneity). The AUC(6.5) (range 0.58 to 0.74, compared to 0.74 under predictor measurement homogeneity) was particularly affected by random predictor measurement heterogeneity. We present results for the Cox regression model under no censoring only. The impact on the measures of predictive performance under administrative and uninformative (random) censoring and for the parametric exponential survival model were similar (data in Online Supplement 1, Section 3).

As measurement procedure W contained more random variability compared to X, i.e., a case of random measurement heterogeneity, $\sigma_\epsilon > 0$, the O/E ratio moved slightly under 1 (Figure 2, row A). The AUC(6.5) and IPA(6.5) decreased as random measurement heterogenei ty increased.

Additive systematic measurement heterogeneity, i.e., $\psi \neq 0$, affected the calibration-in-the-large coefficient at implementation, but minimally affected the AUC(6.5), and IPA(6.5) at implementation (Figure 2, row B). When measurement procedure W at implementation provided a systematically higher value of the predictor compared to

measurement procedure *X* at validation, i.e., $\psi > 0$, this resulted in overestimation of the average outcome incidence at implementation, and the O/E ratio < 1.

Multiplicative systematic measurement heterogeneity, i.e., $\theta \neq 1$, yielded an O/E ratio < 1 in case $\theta > 1$ (Figure 2, row C). Multiplicative systematic measurement heterogeneity minimally affected the AUC(6.5) in absence of additive systematic and random measurement heterogeneity. As θ was further from 1, the IPA(6.5) at implementation decreased, indicating lower overall accuracy.

Combined random, additive systematic, and/or multiplicative systematic predictor measurement heterogeneity sometimes reinforced or cancelled out effects on predictive performance (see Online Supplement 1, Section 3).

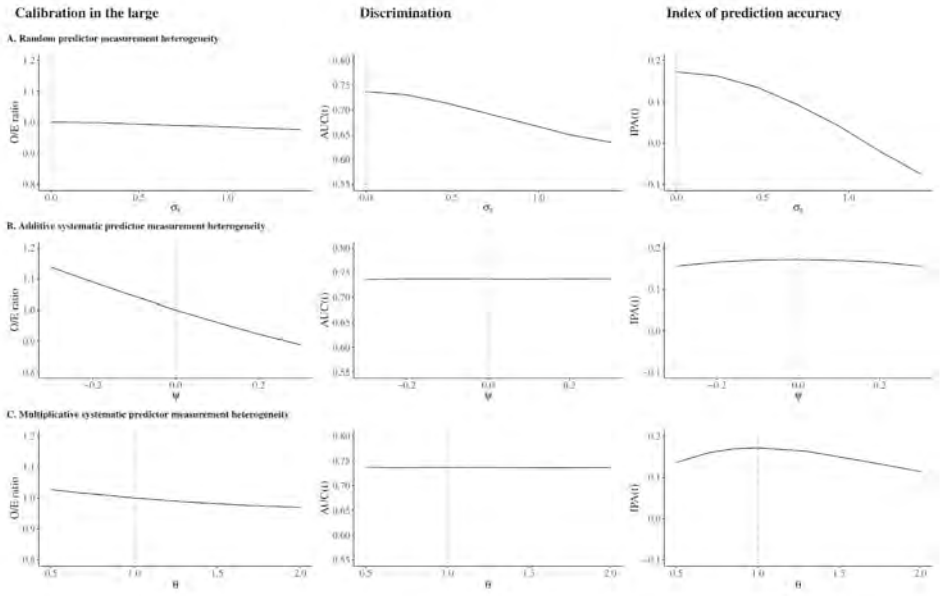


Figure 2. Measures of predictive performance under predictor measurement heterogeneity between validation and implementation setting. Results shown for random predictor measurement only (row A), additive systematic predictor measurement only (row B), and multiplicative systematic predictor measurement heterogeneity only (row C). The vertical dashed line indicates predictor measurement homogeneity between validation and implementation setting. The x-axes show measurement heterogeneity parameters describing the predictor measurement at implementation relative to the predictor measurement at validation, where σ_ϵ denotes random deviations from the measurement at validation, ψ denotes an additive shift with respect to the measurement at validation, and θ denotes a systematic multiplicative association with the measurement at validation. Note that additional simulation scenarios were run to smooth the plots.

4 | Illustration of quantitative prediction analysis

We describe an analysis that quantifies the impact of anticipated predictor measurement heterogeneity between the validation and implementation setting. This is illustrated by validating a prognostic model predicting the 6-year risk of developing type-2 diabetes in a modified example dataset. Section 4.1 describes derivation and validation of the model. The hypothetical step to implementation is described in Section 4.2 by means of the proposed analysis. A detailed description including analysis code can be found in Online Supplement 2.

4.1 | Motivating validation study

Zhang and colleagues derived a prognostic model predicting the 6-year risk of developing type-2 diabetes from the predictors age, BMI, triglyceride, and fasting plasma glucose at moment of prediction[40]. Here, we used the set of predictors identified by Zhang et al. as a starting point for derivation and validation of a prognostic model. We emphasize that this is for illustrative purposes only and is not a recommended approach in practice, where a validation study typically validates a previously derived prognostic model as-is.

The example data was obtained from a publicly available data set containing information about 15,464 individuals who participated in a medical examination program at the Murakami Memorial Hospital from 2004 to 2015, made publicly available alongside a study by Okamura and colleagues[41]. BMI was reported to be measured at medical examination; we assumed it was computed from scale and measuring-tape measurements.

We recreated a derivation and validation sample by resampling from the original dataset stratified on cumulative event fraction at 6 years (2,192 days). In the recreated derivation sample (n = 10,824), the incidence density rate was 2.83/1,000 person years (134 events in total), event times ranged from 285 to 2,191 days, and censoring times ranged from 164 to 2,192 days. In the recreated validation sample (n = 4,639), the incidence density rate was 2.88/1,000 person years (58 events in total), event times ranged from 285 to 2,191 days, and censoring times ranged from 164 to 2,192 days. We assumed censoring was non-informative.

We evaluated predictive performance at 6 years using the performance measures described in our simulation study and used a bootstrap procedure with 500 resamples to correct the AUC(6 years) and IPA(6 years) for optimism and estimate 95-percentile

confidence intervals (CIs). There was no predictor measurement heterogeneity across derivation and validation setting by construction of the samples.

At derivation, the calibration-in-the-large O/E ratio was 1.02 (95% CI, 0.76; 1.43), the optimism-corrected AUC(6 years) was 0.87 (95% CI, 0.84; 0.90), and the optimism-corrected IPA(6 years) was 0.07 (95% CI, 0.04; 0.11). At validation, the calibration-in-the-large O/E ratio was 1.01 (95% CI, 0.78 to 1.34), the AUC(6 years) was 0.89 (95% CI, 0.84 to 0.93), and the IPA(6 years) was 0.06 (95% CI, 0.01 to 0.11).

4.2 | Quantitative prediction analysis for anticipating the impact of predictor measurement heterogeneity between validation and implementation setting on predictive performance

Seven steps are described to perform a quantitative prediction analysis in a prognostic model validation study to assess the impact of anticipated measurement heterogeneity in measurement of BMI, where BMI is assumed to be measured from self-reported height and weight at implementation, instead of tape and scale measures at validation (Box 1).

First, the prediction target is stated. In this example, the prediction target would be the 6-year risk of developing type-2 diabetes in Asian adults presenting for preventive medical examination by measurements of age, BMI, triglyceride, and fasting plasma glucose at moment of prediction. Incident diabetes is defined as HbA1c $\geq$ 6.5% (48 mmol/mol) in two test results, measured using a standardized method[42]. Age is measured in years, BMI is calculated from self-reported weight and height, triglyceride is measured according to standards of the National Institute of Standards and Technology[43], and fasting plasma glucose is measured using a standardized method[44,45]. Details on procedures to measure HbA1c, triglyceride, and fasting plasma glucose are omitted here for brevity, but are ideally described in more detail in an empirical study. Treatment assignment policy was assumed to be similar in the research settings compared to the target clinical setting and interventions such as diet were not modeled explicitly (i.e., ignore-treatment strategy[46]).

Second, it is described whether predictor measurement procedures in the validation setting correspond to those that will be used at implementation. Measurements of age, triglyceride, and fasting plasma glucose roughly correspond to the target predictor measurement procedures. However, the validation study measured BMI during medical examination of a patient, which differs from self-reported measurements defined in the prediction target.

Box 1. Quantitative prediction analysis to quantify the impact of anticipated predictor measurement heterogeneity when implementing a prognostic model in clinical practice (details in Section 4.2 of the main text).

1. State the prediction target.
2. Report whether predictor measurement procedures in the validation setting correspond to those at implementation.
3. Identify one predictor that is expected to be measured using a different procedure in the implementation setting than in the validation setting.
4. Define a model for the relation between the measurement in the validation study and its equivalent in the implementation setting.
5. Perform a literature search to establish a range for the size of the possible parameters of predictor measurement heterogeneity.
6. Simulate the scenarios of anticipated measurement heterogeneity to assess the possible impact on predictive performance.
7. Report the impact of anticipated predictor measurement heterogeneity on predictive performance in clinical implementation.

Third, a predictor is identified that is expected to be measured differently (e.g., using a different procedure) in the implementation setting compared to the validation setting. Measurement heterogeneity was expected to be strongest for the predictor BMI.

Fourth, a model for the relation between the measurement of BMI in the validation study, BMI_{val}, and in the implementation setting, BMI_{imp}, is defined, e.g.:

$$BMI_{imp} = \psi + \theta BMI_{val} + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and $\psi \neq 0$ indicates that measurements of BMI in the implementation setting are systematically additively shifted with respect to BMI in the validation study, $\theta \neq 0$ indicates measurements of BMI in the implementation setting are systematically multiplicatively altered with respect to BMI in the validation study, and $\sigma_\epsilon > 0$ indicates measurements of BMI in the implementation setting contain more random variation relative to BMI in the validation study.

Fifth, the range is specified for the parameter values of the model for the anticipated predictor measurement heterogeneity, as defined in Step 4. A literature search was performed to identify studies describing measurement error in BMI. Informed by studies

comparing measured and self-reported BMI values[47-51], the range of measurement error parameters was specified as -1 to 0 for ψ, 0.9 to 1 for θ, and 0 to 1.5 for σ_ϵ. In general, we advise to use terms like 'measurement error', 'validation study', and the measurement procedures to search for relevant literature. Of note, the term 'validation study' has a different meaning in prediction literature compared to measurement error literature. In prediction literature, a validation study refers to a study that evaluates the predictive performance of an existing prediction model. In measurement error literature, a validation study refers to a study in which a perfect measurement is taken of a mismeasured covariate, usually in a subset of individuals included in the study[52]. The purpose of a measurement-error validation study is to estimate the connection between the error-prone and error-free measurement, for instance using measurement error models, to address issues introduced by measurement error in the substantive analysis. In the current study, we thus far used the term 'validation study' according to the prediction literature.

Sixth, the scenarios of anticipated measurement heterogeneity can be investigated using statistical simulations to assess the possible impact on predictive performance. Briefly, we plugged in the values found in Step 5 into the model specified in Step 4 to generate measurements of BMI that can be anticipated in the implementation setting in participants otherwise similar to the validation sample. We evaluated the O/E ratio for calibration in the large, AUC(6 years), and IPA(6 years) under the scenarios of measurement heterogeneity in BMI (see Online Supplement 2) and plotted the outcomes (Figure 3).

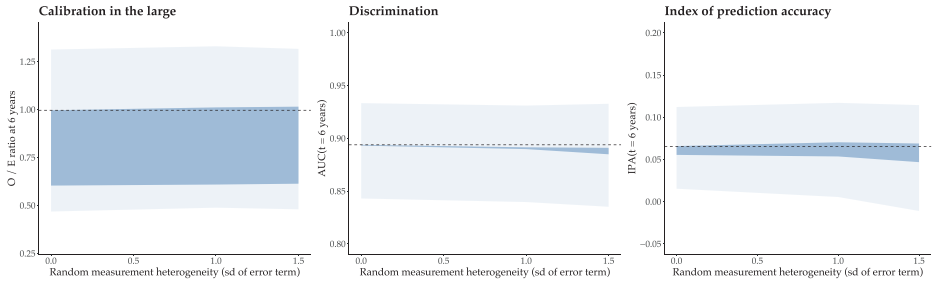


Figure 3. Impact of anticipated heterogeneity in measurement of the predictor body mass index on measures of predictive performance at implementation of a model to predict the 6-year risk of developing diabetes type 2. Dark blue indicates the impact within the range of specified predictor measurement heterogeneity and light blue indicates 95 percentile CIs from 500 bootstrap resamples. Random predictor measurement heterogeneity is presented on the x-axis, and performance measures are marginalized over scenarios of additive and multiplicative systematic predictor measurement heterogeneity.

Seventh, the impact of anticipated predictor measurement heterogeneity on predictive performance in the implementation setting can be reported in a validation study, accompanied by a description of Steps 1-6. Anticipating on the possibility that BMI may be measured differently in clinical practice compared to how data on BMI were collected in the validation study, we found that performance of the type-2 diabetes prediction model might be reduced when implemented 'as-is' in clinical practice (Figure 3). In particular, with increasing differences in BMI measurement variance between our validation study and the clinical target setting, model miscalibration increases. Possible consequences of this finding may be to either update the current prediction model using self-reported measures of BMI before implementing it in clinical practice or to collect data on BMI using scale and measuring-tape measures only when the model is used in clinical practice to predict 6-year risk of developing diabetes.

5 | Discussion

Our simulations indicated that predictor measurement heterogeneity across the validation and implementation setting of a prognostic model can substantially affect predictive performance at implementation. We illustrated how a quantitative prediction analysis can be applied in validation studies to quantify the impact of anticipated dissimilar predictor measurements in the clinical target setting on predictive performance. Based on this analysis, a validation study can inform readers about the severity of possible predictor measurement heterogeneity when the model is implemented in clinical practice.

The rationale for the quantitative prediction analysis was analogous to the quantitative bias analysis framework by Lash and colleagues, which can be applied to estimate the direction, magnitude, and uncertainty from systematic errors affecting studies of causal inference[53,54]. While Lash and colleagues encourage researchers to address multiple sources of bias[53], we focused on a single source of heterogeneity across settings that can affect performance of a clinical prediction model. In this, we focused on non-differential systematic and random measurement heterogeneity in a single predictor, where the clinical implementation setting contained more measurement variance compared to the validation setting. Future work could extend these quantitative prediction analyses to non-differential measurement heterogeneity, to settings were the clinical implementation setting contained less measurement variance compared to the validation setting – for instance through methods analogous

to the simulation-extrapolation method (SIMEX)[55,56] – and to models that take into account correlations of measurement heterogeneity structures when multiple predictors are expected to be measured heterogeneously across validation and implementation setting. Additionally, other sources of heterogeneity across settings that can affect performance of a clinical prediction model can be added to the quantitative prediction analysis, such as heterogeneity in event rate, heterogeneity in outcome measurement procedures, and heterogeneity in treatment-assignment policies during follow-up.

The example of predicting the risk of developing type-2 diabetes illustrated the impact of anticipated measurement heterogeneity in the predictor BMI. Notably, the magnitude of the impact of anticipated measurement heterogeneity strongly depends on whether the linear predictor was centered to the validation data. While many functionalities in R[39] center the linear predictor by default, centering is likely uncommon in clinical practice and obviously decreases the impact of predictor measurement heterogeneity on predictive performance. A limitation of the example is that only measurement heterogeneity in a single predictor was considered, while the predictor fasting plasma glucose can potentially be measured heterogeneous across settings as well, in particular because fasting instructions and adherence to instructions may differ across settings. Taking this into account requires consideration of the duration of fasting relative to the timing of the plasma glucose measurement[57]. Modelling the functional form of fasting plasma glucose or another (circadian) fluctuating hormone or biomarker over time to assess the impact in heterogeneity of measurement timings across time would be an interesting topic for future research.

As a limitation to our study, implementation of the quantitative prediction analysis may be hampered because literature informing the choice of measurement error parameters (Step 5) may be limited. When no information is available about predictor measurement structures in an implementation setting of interest, it might be helpful to set up a (measurement heterogeneity) validation study to estimate the predictor measurement heterogeneity parameters directly[52]. This may be an alternative approach to anticipate the performance of a prognostic model in a particular setting that is likely less cumbersome than conducting a prediction validation study in the implementation setting.

Data for derivation and validation of prognostic models are collected ideally using procedures that match the target clinical setting. When this is infeasible, the quantitative prediction analysis provides an analytical approach to quantify the anticipated impact of the discrepancies between available research data and clinical practice.

Online Supplementary Files

The supplementary files referred to in this Chapter are available online at https://github.com/KLuijken/Dissertation_Online_Supplements/tree/main/Chapter_8

References

1. Steyerberg EW. *Clinical Prediction models*. Springer; 2019.
2. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine*. 1999;130(6):515-524.
3. Shmueli G, Koppius OR. Predictive analytics in information systems research. *MIS Quarterly*. 2011;35(3):553-572.
4. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *British Medical Journal*. 2013;346.
5. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381.
6. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11(10):e1001744.
7. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine*. 2015;162(1):W1-W73.
8. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine*. 2013;32(18):3158-3180.
9. Steyerberg EW, Uno H, Ioannidis JP, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. *Journal of Clinical Epidemiology*. 2018;98:133-143.
10. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in Medicine*. 2016;35(2):214-226.
11. Pajouheshnia R, Van Smeden M, Peelen L, Groenwold R. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *Journal of Clinical Epidemiology*. 2019;105:136-141.
12. Luijken K, Groenwold RH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. *Statistics in Medicine*. 2019;38(18):3444-3459.
13. Khudyakov P, Gorfine M, Zucker D, Spiegelman D. The impact of covariate measurement error on risk prediction. *Statistics in Medicine*. 2015;34(15):2353-2367.
14. Rosella LC, Corey P, Stukel TA, Mustard C, Hux J, Manuel DG. The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model. *Population Health Metrics*. 2012;10(1):1-11.
15. Luijken K, Wynants L, van Smeden M, et al. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *Journal of Clinical Epidemiology*. 2020;119:7-18.
16. Wynants L, Collins GS, Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG: An International Journal of Obstetrics & Gynaecology*. 2017;124(3):423-432.
17. Cowley LE, Farewell DM, Maguire S, Kemp AM. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagnostic and Prognostic Research*. 2019;3(1):1-23.
18. Toll D, Janssen K, Vergouwe Y, Moons K. Validation, updating and impact of clinical prediction rules: a review. *Journal of Clinical Epidemiology*. 2008;61(11):1085-1094.
19. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. *British Medical Journal*. 2009;338.
20. Riley RD, Ensor J, Snell KI, et al. Calculating the sample size required for developing a clinical prediction model. *British Medical Journal*. 2020;368.

21. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine.* 1996;15(4):361-387.

22. Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans M, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology.* 2001;54(8):774-781.

23. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *British Medical Journal.* 2009;338.

24. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine.* 2000;19(4):453-473.

25. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *British Medical Journal.* 2009;338.

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

27. Ensor J, Snell KI, Debray TP, et al. Individual participant data meta-analysis for external validation, recalibration, and updating of a flexible parametric prognostic model. *Statistics in Medicine.* 2021;40(13):3066-3084.

28. Adams ST, Leveson SH. Clinical prediction rules. *British Medical Journal.* 2012;344.

29. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of Internal Medicine.* 2006;144(3):201-209.

30. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine.* 2019;38(11):2074-2102.

31. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine.* 2005;24(11):1713-1723.

32. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology.* 2016;74:167-176.

33. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Medicine.* 2019;17(1):1-7.

34. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics.* 2005;61(1):92-105.

35. Uno H, Cai T, Tian L, Wei L-J. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association.* 2007;102(478):527-537.

36. Blanche P, Kattan MW, Gerds TA. The c-index is not proper for the evaluation of-year predicted risks. *Biostatistics.* 2019;20(2):347-357.

37. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review.* 1950;78(1):1-3.

38. Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and Prognostic Research.* 2018;2(1):1-7.

39. R Core Team. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2020.

40. Zhang M, Zhang H, Wang C, et al. Development and validation of a risk-score model for type 2 diabetes: a cohort study of a rural adult Chinese population. *PloS one.* 2016;11(4):e0152054.

41. Okamura T, Hashimoto Y, Hamaguchi M, Obora A, Kojima T, Fukui M. Ectopic fat obesity presents the greatest risk for incident type 2 diabetes: a population-based longitudinal study. *International Journal of Obesity.* 2019;43(1):139-148.

42. American Diabetes Association. 2. Classification and diagnosis of diabetes: Standards of Medical Care in Diabetes—2021. *Diabetes Care.* 2021;44(Supplement 1):S15-S33.

43. Warnick GR, Kimberly MM, Waymack PP, Leary ET, Myers GL. Standardization of measurements for cholesterol, triglycerides, and major lipoproteins. *Laboratory Medicine.* 2008;39(8):481-490.

44. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation. 2006.

45. D'Orazio P, Burnett RW, Fogh-Andersen N, et al. Approved IFCC recommendation on reporting results for blood glucose: International Federation of Clinical Chemistry and Laboratory Medicine Scientific Division, Working group on selective electrodes and point-of-care testing (IFCC-SD-WG-SEPOCT). *Clinical Chemistry and Laboratory Medicine (CCLM).* 2006;44(12):1486-1490.

46. van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *European Journal of Epidemiology.* 2020;35:619-630.

47. Nawaz H, Chan W, Abdulrahman M, Larson D, Katz DL. Self-reported weight and height: implications for obesity research. *American Journal of Preventive Medicine.* 2001;20(4):294-298.

48. Allison C, Colby S, Opoku-Acheampong A, et al. Accuracy of self-reported BMI using objective measurement in high school students. *Journal of Nutritional Science.* 2020;9:e35.

49. Dekkers JC, van Wier MF, Hendriksen IJ, Twisk JW, van Mechelen W. Accuracy of self-reported body weight, height and waist circumference in a Dutch overweight working population. *BMC Medical Research Methodology.* 2008;8(1):1-13.

50. Villarini M, Acito M, Gianfredi V, et al. Validation of self-reported anthropometric measures and body mass index in a subcohort of the dianaweb population study. *Clinical Breast Cancer.* 2019;19(4):e511-e518.

51. Ortiz-Panozo E, Yunes-Díaz E, Lajous M, Romieu I, Monge A, López-Ridaura R. Validity of self-reported anthropometry in adult Mexican women. *Salud publica de Mexico.* 2017;59:266-275.

52. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement error in nonlinear models: a modern perspective.* Chapman and Hall/CRC; 2006.

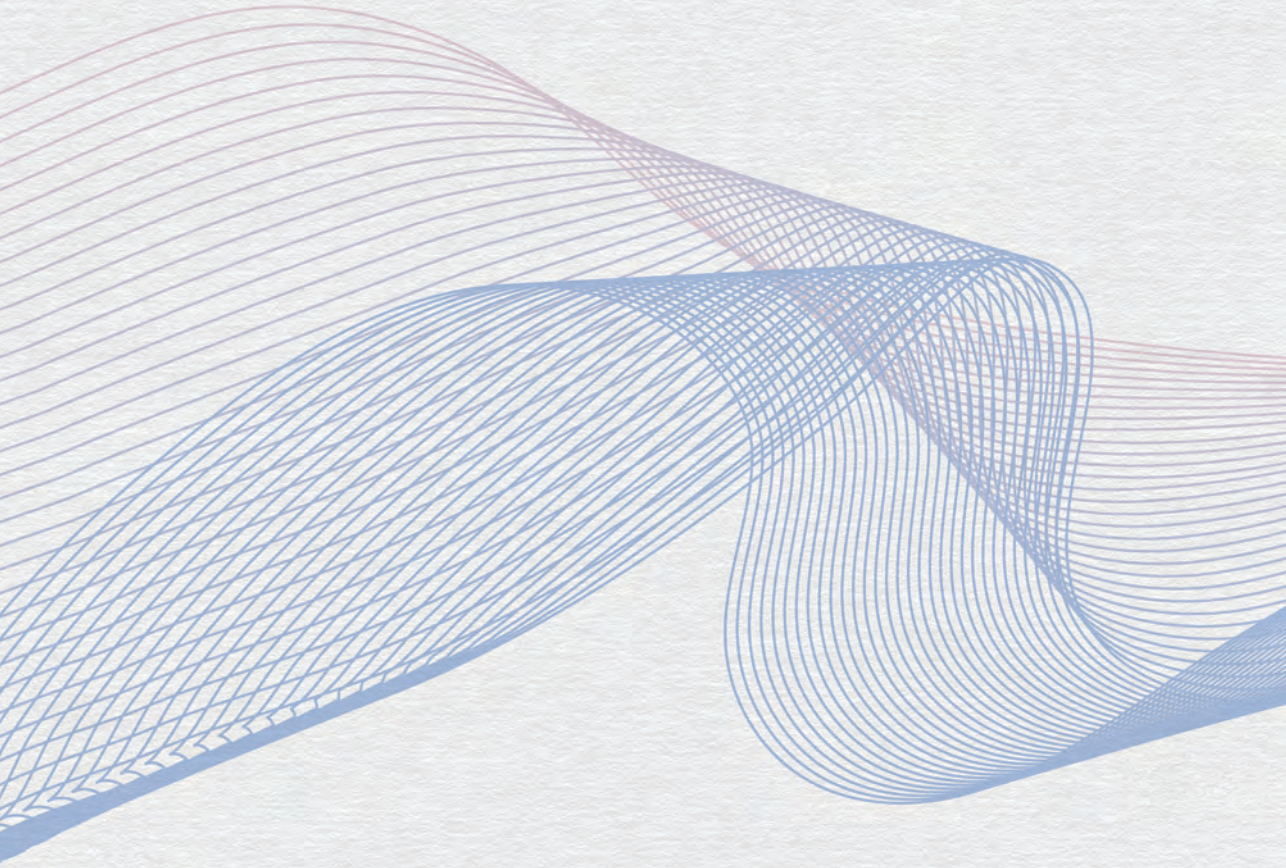
53. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *International Journal of Epidemiology.* 2014;43(6):1969-1985.

54. Lash TL, Fox MP, Fink AK. *Applying quantitative bias analysis to epidemiologic data.* Springer Science & Business Media; 2011.

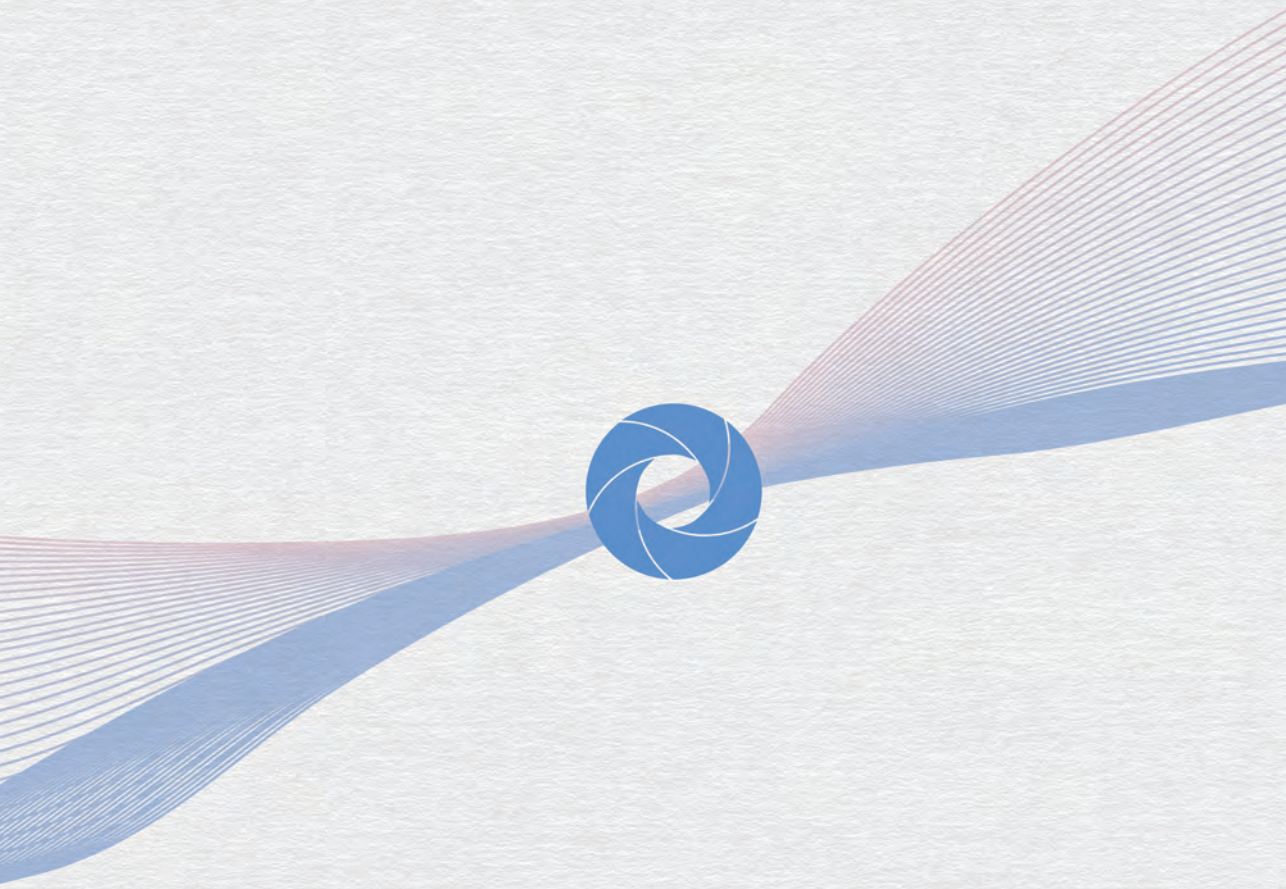
55. Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association.* 1994;89(428):1314-1328.

56. Stefanski LA, Cook JR. Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association.* 1995;90(432):1247-1256.

57. Whittle R, Royle K-L, Jordan KP, Riley RD, Mallen CD, Peat G. Prognosis research ideally should measure time-varying predictors at their intended moment of use. *Diagnostic and Prognostic Research.* 2017;1(1):1-9.



9



Summary and General discussion

Summary

Over the last decades, epidemiological methods have been refined, increasingly so in the last years, making it challenging to keep abreast of all methodological developments. The choice of the data analytical method directly influences the interpretation and clinical meaning of results of an analysis, yet it is undesirable that technical considerations define the subject of the investigation. Having a deeper understanding of the impact that data analytical decisions can have on the interpretation of numerical results of a study would help to apply analytical tools that are both suitable and appropriate to answer clinical questions. The aim of this thesis was to investigate the impact of choices regarding the design and statistical analysis of a study on the meaning of its numerical results in two sets of case studies in research into causal effects (Part I) and prediction research (Part II). The main findings of the investigation are summarized below.

Part I: Impact of applied methods on the meaning of numeric estimates in studies of causal effects

The studies described in Part I of this thesis provided examples of the impact of choices regarding the design and statistical analysis on numerical results in studies of causal effects. This part outlined the many data analytical decisions to be made in those studies. Chapters 2 and 4 highlighted two particular decisions and indicated that the interpretation of effect estimates in studies of causal effects critically depends on the choice of the study time origin and covariate selection strategy. Results of these studies imply that it is important to avoid a backward process of implicitly letting the applied study design and statistical analysis define the meaning of the study results. This is underlined by the studies described in Chapters 3 and 5, which discussed how a clearly formulated study aim can clarify whether clinical interest, research conduct, and interpretation of results are appropriately aligned. However, stating the study aim is challenging in clinical studies, which is reflected in the low frequency of explicitly reported estimands, i.e., the quantity of interest that answers (or best approximates) the clinical research question, in studies of pharmacological treatment effects found in Chapter 2.

Chapter 2 described that the impact of operationalization of the study time origin on numerical results was difficult to assess in pharmacoepidemiologic studies because of incomplete reporting. The reporting on choices regarding operationalization of the time

origin of a research design was investigated in a review of 89 comparative effectiveness and safety cohort studies published in six high-ranked pharmacoepidemiologic journals in 2018 and 2019. Forty percent of studies reported implementing a new-user design and 13% reported implementing a prevalent-user design. Alignment of start of follow-up, moment of meeting eligibility criteria, and treatment initiation was reported to be aligned in 53% of studies implementing a new-user design (19 out of 36 studies) and was insufficiently described in 42% of studies implementing a new-user design. The validity of the operationalization of the study time origin can only be assessed with respect to the study estimand. However, the estimand was explicitly reported in only 22% of studies implementing a new-user design.

In **Chapter 3**, the rigor with which the study aim is defined in exploratory etiologic studies was linked to the interpretation of findings of those studies. A continuum of scrutiny for study conduct was defined, ranging from ad-hoc to targeted. Where an exploratory etiologic analysis is situated on this continuum directly affects interpretability of findings. We argued that acting upon results from ad-hoc analyses as if they rose from targeted analyses by performing further confirmatory studies or by implementing them in clinical practice can contribute to research waste and might harm patients. Practical pointers for good practice in exploratory etiological research were provided, such as the use of rigorous methodologic and statistical approaches and taking responsibility for exploratory findings by reporting a clear agenda for future research.

The study described in **Chapter 4** illustrated that applying backward elimination to reduce the set of covariates for confounding adjustment was rarely more efficient than covariate selection based on causal knowledge. An expression was derived that quantifies how variable omission relates to bias and variance of effect estimators. Simulations were conducted to investigate if and under which conditions causal effect estimation in observational studies can improve by using backward elimination on a prespecified set of potential confounders. Applying backward elimination did not reduce the mean squared error of effect estimators compared to a full model including all prespecified covariates, yet bias was increased. In less than 3% of the 3,960 scenarios considered, the mean squared error of effect estimators was slightly lower when backward elimination was used compared to the full model. Hence, when an initial set of potential confounders can be specified based on background knowledge, our findings indicated there is minimal added value of backward elimination.

In **Chapter 5** an assessment was proposed regarding choices in the design and statistical analysis of studies included in systematic reviews of operative interventions. Intended as a first proposal for summarizing key information needed to assess applicability and methodological quality of studies, we derived an easy-to-use set for initial evaluation of studies of operative interventions based on existing risk of bias tools. The set contained nine items: population, intervention, comparator, outcome, confounding, missing data and selection bias, intervention status, outcome assessment, and pre-specification of analysis. The assessment of applicability and methodological quality can be done as part of a systematic review to discard studies of low quality with relative ease and to separate out higher quality studies for further scrutiny of methodological quality using available assessment tools.

Part II: Impact of applied methods on the meaning of numeric estimates in prediction modelling studies

The studies described in Part II of this thesis focused on the impact of changes in predictor measurement strategies across settings on performance of prediction models. Such changes are referred to as predictor measurement heterogeneity. The phenomenon predictor measurement heterogeneity was formally defined using measurement error models, which allowed for an investigation of the implications of predictor measurement heterogeneity through analytical approaches, simulations, analysis of empirical datasets, and a proposed quantitative prediction analysis. All of these indicated that even when all other factors, such as the modelling strategy, outcome prevalence, included predictors, and patient characteristics, were constant across settings, a change in measurement procedure affected the performance of prediction models. This fosters reconsideration of the way prediction models are specified, and particularly whether predictor measurement procedures should be an integral part of the model specification.

Chapter 6 described how predictor measurements are linked to clinical applicability of predictions of binary logistic prediction models using analytical and simulation approaches. An established taxonomy of measurement error models was used to define and clarify the phenomenon called predictor measurement heterogeneity: variation in the measurement of predictor(s) between the derivation of a prediction model and its validation or implementation. Using analytical and simulation approaches, it was shown that out-of-sample performance of binary logistic prediction models can be hampered

when predictors are measured differently at derivation and external validation. These findings highlight that it is insufficient to describe a prediction target in general terms without specifying the procedures with which predictors are (to be) measured.

In **Chapter 7** it was shown how predictor measurements are linked to clinical applicability of predictions of binary logistic diagnostic models using empirical illustrations in three clinical datasets. Nine scenarios of predictor measurement heterogeneity were evaluated in previously developed prediction models for diagnosis of ovarian cancer, mutation carriers for Lynch syndrome, and intrauterine pregnancy. Changing the measurement procedure of a predictor influenced the performance at validation of the diagnostic models, most notably model calibration, with the calibration-in-the-large coefficient ranging -0.70 to 1.43 and the calibration slope ranging from 0.50 to 1.67 at validation.

Chapter 8 described a quantitative prediction analysis to anticipate the impact of changes in predictor measurement strategies for prognostic time-to-event models. Using simulations with various scenarios of predictor measurement heterogeneity, we showed that out-of-sample performance can be hampered when predictors are measured differently at validation and implementation for time-to-event outcome models. A quantitative prediction analysis was proposed to quantify the impact of anticipated predictor measurement heterogeneity across validation and implementation setting.

General discussion

Each chapter of this thesis characterized one or multiple data analytical decisions and described the impact they might have on the interpretation of numerical results. In Chapters 4 and 6, the impact of data analytical decisions was first studied using relatively simple analytical expressions, followed by simulation studies examining implications of the data analytical decision that could not be described using closed-form solutions. The combination of these approaches provided complementary insights contributing to the aim of this thesis. In Chapter 4, the analytical expression specified how omitting a variable from the data analytical model relates to bias and variance of effect estimators. Based on this result, scenarios can be defined in which covariate omission is beneficial in theory, but the simulations indicated that this benefit rarely occurred in settings more realistic for clinical studies where covariates are automatically selected rather than omitted. Since some methodological papers pointed out that backward elimination could be applied for selection of potential confounders[1-3], the simulation findings shed light on the frequency and type of situations in which this might be considered beyond theoretical considerations.

In Chapter 6, analytical expressions specified how measurement error in predictors relates to within-sample discrimination and overall accuracy of binary logistic prediction models. Subsequently, the taxonomy of measurement error models was used to define predictor measurement heterogeneity across settings of derivation, validation, and implementation. Simulation studies were conducted to quantify the impact of predictor measurement heterogeneity on predictive performance at external validation. The formal definition of predictor measurement heterogeneity and simulation findings added to existing literature, which thus far described the importance of the choice of predictor measurement[4,5] and impact of measurement error on within-sample predictive performance[6,7]. Predictor measurement heterogeneity models can help to further explain discrepancies in predictive performance between settings. Researchers can use these models to quantify the impact of anticipated predictor measurement heterogeneity in empirical studies, as was explained in Chapter 8.

Another approach taken in this thesis to understand how the clinical interpretation of estimates depends on data analytical decisions was to motivate the methodological investigation by clinical interests. This was done by examining published clinical studies with a team involving at least one practicing clinician (Chapter 2, 3, and 5) and by presenting a motivating clinical example and analyzing the (modified) empirical

data (Chapter 4 and 7 - 8). As a limitation, the focus was on the implication of applying a *method*, and not on zooming in on the *clinical* meaning of estimates resulting from that particular method. The interpretation of numerical results could have been described more deeply in an empirical study driven by a clinical research question.

In general, we have outlined the impact that various data analytical decisions can have on the meaning of estimates using different approaches. However, in hindsight, the investigations were essentially conducted in reverse order. The chapters describe the impact of a technique on the results, while ideally the desired meaning of the result dictates the choice of methods. An important limitation is therefore that we have not investigated the impact of defining the desired result, i.e., specifying the estimand, on the choice of methods. This relevant topic should receive more attention in future research. The remainder of this chapter will therefore discuss directions for future research into defining estimands in clinical studies.

Recommendations for future investigations into targeted research questions

When the research question of a study is posed in generic terms, this leaves room for a mismatch between the interpretation of the estimate produced by the applied design and statistical analysis and the quantity of clinical interest. Clearly defining the estimand could be the missing link to prevent such disparities (Figure 1). Yet, as the findings of Chapter 2 indicated, formulating an estimand is challenging.

Research has been conducted and is ongoing to define clinical research questions such that they can guide a quantitative analysis. The importance of defining an estimand was bolstered by ICH E9(R1) addenda on estimands published since 2010[8] and further guidance to put ICH E9(R1) into practice when conducting a randomized clinical trial was given in 2020[9-11]. For observational studies of causal inference, the concept of 'sufficiently well-defined interventions' has been introduced to formulate research questions such that numerical effect estimates can be causally interpreted[12-19]. Additionally, to avoid having to state mathematical expressions such as the distribution of potential outcomes, the principle of 'target trial emulation' has been introduced for explicating the estimand of a study[20].

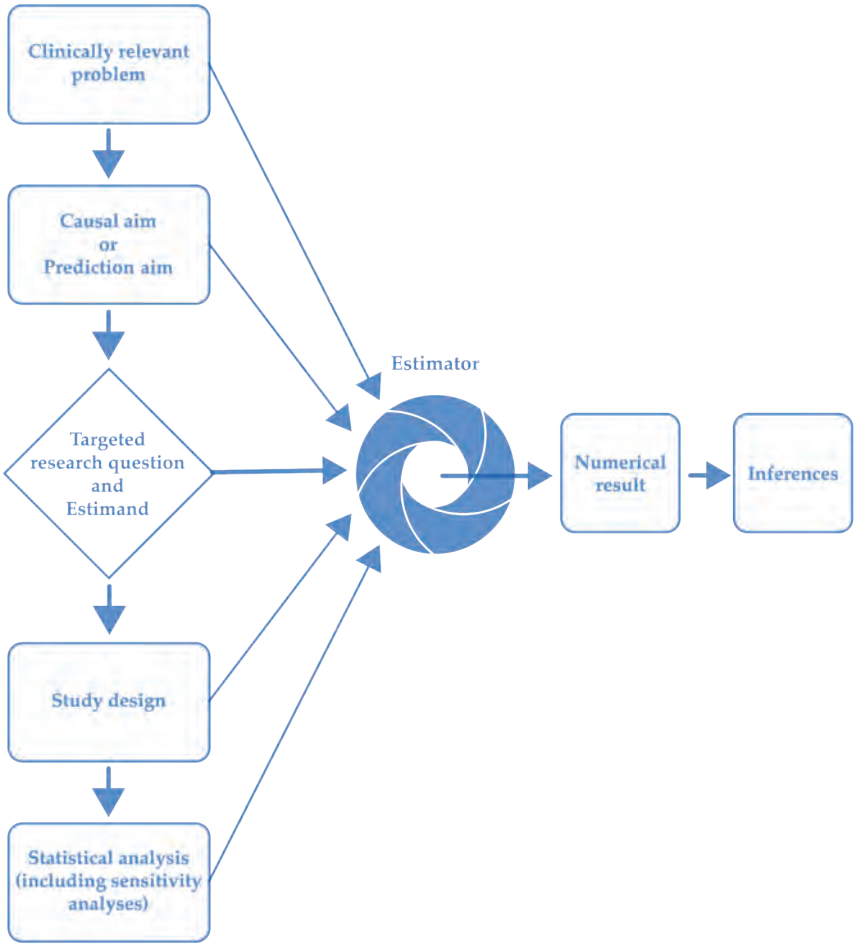


Figure 1. Adding a missing link to the schematic depiction of stages of research conduct of a quantitative study. When data analytical decisions are not driven by substantial clinical considerations, this could result in a disparity between the clinical research question and the meaning of a numerical result. A targeted research question and estimand could be the missing link to prevent such mismatches, by guiding the choice of (statistical) methods that yield a numerical result with the desired interpretation.

Reasons why estimands are not always explicitly defined in clinical studies could be that they require a profound understanding of statistical theory and that few recommendations are available for non-therapeutic research. Further guidance could thus be developed to help researchers arrive at a sufficiently well-defined estimand, starting out from a clinical perspective, i.e., by defining a 'targeted research question'. We suggest five topics for future research that can contribute to making targeted research questions more central to clinical research (Box 1).

Box 1 Five suggested topics for future research on targeted research questions.

1. Distinguish a clinical perspective from a statistical perspective with the aim of aligning expert input
2. Identify the optimal balance between succinctness and completeness of targeted research questions
3. Specify how and when targeted research questions differ between studies of causal inference and prediction modelling studies
4. Learn about targeted research questions through empirical exemplar studies
5. Teach formulating targeted research questions

1. Distinguish a clinical perspective from a statistical perspective with the aim of aligning expert input

Performing clinical research requires a different mindset than the process of clinical reasoning. Loosely described, clinical reasoning is an iterative process of integrating clinical knowledge with patient information to support a final diagnosis and medical decisions[21]. It serves the physician to start out with a broad differential diagnosis, to briefly probe several hypothesized diagnoses, and to be wary of not deciding on a diagnosis until (s)he has interviewed and observed the patient, as cognitive errors might then interfere with the perception of the problem[22]. Keeping an open view helps to make the most auspicious clinical decisions.

The general order of reasoning in statistics appears to be the reverse. Loosely described, statistical reasoning enables interpretation of empirical data through a process of connecting mathematical arguments with observed phenomena. It serves the statistician to work systematically through phases of making assumptions about a data-generating mechanisms, identifying an estimand and then finding a suitable estimator given the properties of the available data and sampling characteristics of the

estimator – all prior to observing the data[23]. Keeping a principled view helps ensuring that the analysis is mathematically valid.

An important tool for defining the statistical estimation problem is a statistical model. A succinct overview of the perks and perils of statistical models is given at the first pages of the introductory statistics book *Statistical Rethinking*, by McElreath. Statistical models are compared to a kabbalistic Golem: a clay robot that is "animated by truth, but lacking free will, [and doing] exactly what it is told"[24, p. 1]. McElreath explains that statistical models are similar: "Rather than idealized angels of reason, scientific models are powerful clay robots without intent of their own [...] [A scientific model] doesn't discern when the context is inappropriate for its answers. It just knows its own procedure, nothing else. It just does as it's told"[24, p. 2]. To appropriately "tell" a statistical model what aim to serve, the clinical research question must be translated to a quantity of interest, i.e., an estimand. Using statistical models as part of statistical reasoning yields estimates with a clear interpretation, but the process of generating them may be too rigid or nitty-gritty to inform clinical reality. From the perspective of clinical reasoning however, it is clear what information would be a relevant finding, but the scope and complexity of the desired result may be broader than a statistical model can address.

Efforts to put estimands at the center of clinical research may be better received if researchers are more aware of the perspective they naturally take on (be it clinical or statistical) and the strengths and pitfalls of that view. Specifically, it would be valuable to better understand how clinicians can adopt a research perspective *while* contributing their clinical expertise, which is (by definition) vital to clinical research. An overview of the two perspectives and their complementary contributions to clinical research would be helpful in this regard. Special attention could be given to the role of targeted research questions as they inform how to design a study and statistical analysis such that the most applicable clinical evidence can be generated in a language understandable by clinicians and statisticians.

2. Identify the optimal balance between succinctness and completeness of targeted research questions

For a targeted research question to guide study conduct, it must be clear from its specification how to design the study. Obvious as this may sound, it is challenging to formulate a complete yet concise research problem. This is further complicated by the fact that discoveries made *during* a research project may influence which topic is of

interest and thus change the research question. To accommodate iterative refinement of a research question, it seems relevant to develop methods that evaluate whether a research question has sufficient 'targeting' capacity.

A potential development direction for such a method could be a checklist intended to assist clinical investigators with understanding the purpose of their research from a clinical point of view, rather than from quantitative considerations. Such a checklist ideally contains questions rather than criteria. The intended way of use would be to answer a set of questions multiple times throughout a research project, mostly at the start of the project, to have focused answers established before consulting a statistician to support decisions regarding the design and analysis of the study, and to go over the questions once more when the results are obtained. Because the items are stated as questions, the checklist can be thought of as a more formative evaluation that helps cultivating a critical attitude towards the rigor of the study aim, rather than a summative checklist that states what a project should ultimately contain[25], which is more common to protocol, risk of bias, or reporting checklists[26]. Examples of topics that might be addressed include questions that help to identify the overall objective of the study, to target a research question prior to data analysis, to further target the research question at later stages of the analysis, and to report the established targeted research question.

3. Specify how and when targeted research questions differ between studies of causal inference and prediction modelling studies

An extensive body of literature seems to have established which elements specify a well-defined targeted research question and estimand in therapeutic studies of pharmacological interventions[1]. Yet, defining these elements in therapeutic research of complex interventions is arguably less straightforward[27]. For instance, in studies of operative interventions there is a clear variation in relative importance of elements of the complex intervention under study. Although the intervention strategy technically consists of a combined operative (point) intervention and postoperative (longitudinal) treatment regimen, the effect of interest is generally that of the operative intervention. Implicitly this might be acknowledged by ignoring postoperative treatment (invoking

1 Being the target population, treatment strategies being compared (with five main treatment strategies; treatment policy strategy, composite strategy, while-on-treatment strategy, hypothetical strategy, and principal stratification), treatment assignment procedures, follow-up period, outcome of interest (what and when), and causal contrast(s) of interest.

the assumption that it is either irrelevant to the effect of interest or similar across the target population, making the effect found to be at least generalizable), but the choice of estimand is, again, preferably explicit. Defining targeted research questions and estimands that explicitly consider relative importance within complex interventions seems an important item on the agenda of future research into operative interventions.

Specifying a naturally occurring exposure such that it is sufficiently well-defined is arguably less straightforward than specifying assigned interventions[13,28]. The question how to capture the causal impulse of a phenomenon is therefore an important area of future work on etiologic targeted research questions. Investigations that might be helpful in this regard include studies on how exposure time origins should be anchored (relative to calendar time or other events) and whether exposure trajectories could be mapped to treatment strategies as defined by the ICH[8], including a discussion of deviations and approaches to address them. It is not unlikely that these considerations depend so heavily on the clinical context of a specific etiological study that such research should be conducted within a particular clinical field (see also the section on exemplars below).

The specification of estimands has received less explicit attention in the context of prediction research. Important elements of the research question have been defined[4] and discussed as separate topics such as defining the timing of the prediction[29] and addressing intercurrent events similar to the ICH E9(R1) strategies[30,31]. However, to our knowledge, only one unpublished manuscript explicitly discusses how to define a prediction target in prediction research[32]. Further research on prediction targets is particularly eminent because prediction targets do not map to causal estimands one-to-one. A causal estimand expresses a hypothetical effect in a target population and the term "estimand", i.e., "that which is to be estimated", directly refers to this effect. In prediction research, the quantity that is targeted is a conditional probability of the outcome of interest, and it is up for discussion whether an abstract quantity can be pinpointed that represents the corresponding targeted conditional probability[33].

4. Learn about targeted research questions through empirical exemplar studies

The recommendations for future research stated so far mainly considered theoretical developments. However, it is likely that these insights can also be acquired by conducting clinical research according to best known practice. One example of such research is a study on the effect of zidovudine on the survival of human immune-

deficiency virus-positive men[34]. Lessons from such *exemplar* studies might make important contributions to methodological research.

5. Teach formulating targeted research questions

As a final recommendation, targeted research questions merit a more central position in teaching on clinical research for both students with a clinical background and with a statistical background to emphasize that numbers do not speak for themselves.

Teaching material for students with a clinical background could explain how a study design and analysis can be aligned with a certain aim. Some textbooks already head in this direction, e.g.,[35-37], but the guiding principle of the estimand can be put more at the forefront, similar to e.g. Van der Laan and Rose[38] (yet aimed at a less technical audience). A possible corresponding teaching structure is to assign reproducibility projects rather than having students conduct a research project from scratch. Especially for researchers new to the field, questioning decisions made in an existing study might be a fruitful learning strategy. Not having to consider the myriad decisions that need to be made when performing a research project (and most likely being deliberately discouraged from reaching the level of expertise needed to make informed choices) likely results in more mind space to focus on the (mis)alignment in study aim, operationalization, and interpretation.

Teaching material for students with a statistical background could clarify more systematically how a technical procedure follows from the overall aim of the study. This would fit with the current new wave of statistics education that no longer starts with formal probability theory as a basis for statistical inference and emphasizes non-technical issues such as the importance of the research question and data quality[39-42]. Ideally, each vignette or technical assignment contains a reminder of the purpose for which the procedure is applied and questions the alignment and interpretation with respect to that aim.

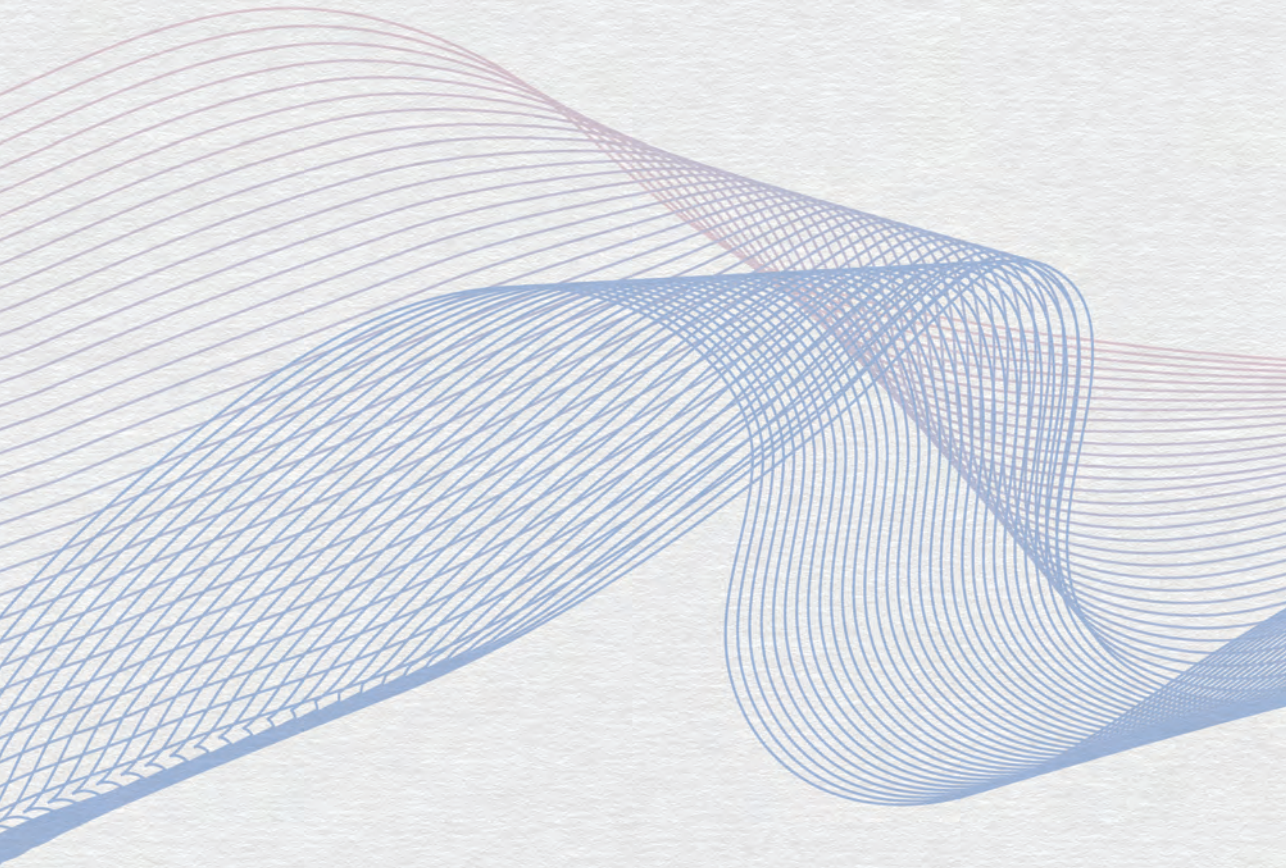
In conclusion

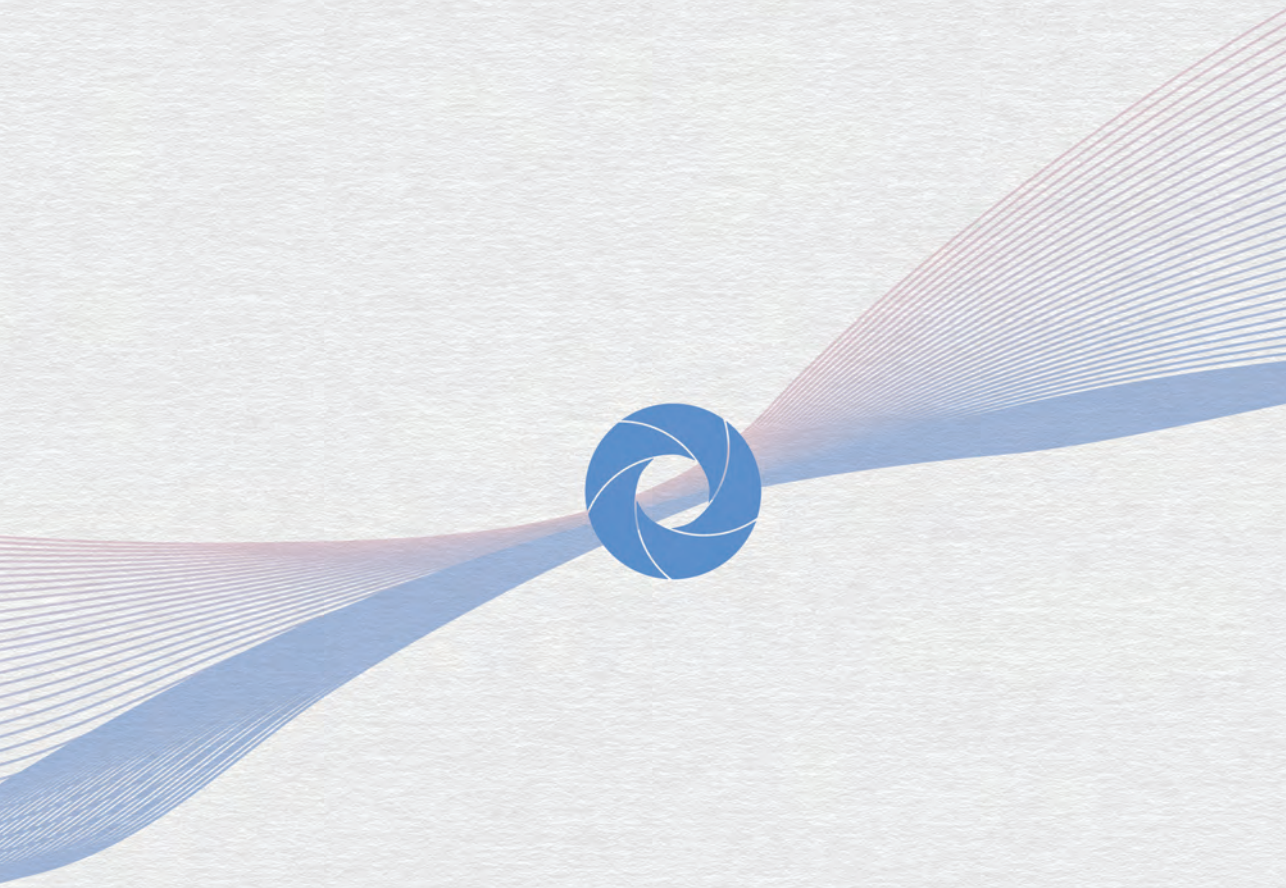
This thesis described the impact of various choices regarding the design and statistical analysis of a study on the meaning of its numerical results. Clearly defining a clinically relevant estimand ensures that data analytical decisions yield meaningful results. Making targeted research questions central to quantitative clinical research can reduce fallacious confidence in (complex) methods and can add to intelligibility of findings.

References

1. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics.* 2011;67(4):1406-1413.
2. Greenland S, Daniel R, Pearce N. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *International Journal of Epidemiology.* 2016;45(2):565-575.
3. Dunkler D, Plischke M, Leffondré K, Heinze G. Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. *PloS one.* 2014;9(11):e113677.
4. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine.* 2015;162(1):W1-W73.
5. Wolff RF, Moons KG, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine.* 2019;170(1):51-58.
6. Khudyakov P, Gorfine M, Zucker D, Spiegelman D. The impact of covariate measurement error on risk prediction. *Statistics in Medicine.* 2015;34(15):2353-2367.
7. Rosella LC, Corey P, Stukel TA, Mustard C, Hux J, Manuel DG. The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model. *Population Health Metrics.* 2012;10(1):1-11.
8. ICH E9 working group. ICH E9 (R1): addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Published 2020. Accessed 28-05-2021.
9. Ratitch B, Bell J, Mallinckrodt C, et al. Choosing estimands in clinical trials: putting the ICH E9 (R1) into practice. *Therapeutic Innovation & Regulatory Science.* 2020;54(2):324-341.
10. Mallinckrodt C, Bell J, Liu G, et al. Aligning estimators with estimands in clinical trials: putting the ICH E9 (R1) guidelines into practice. *Therapeutic Innovation & Regulatory Science.* 2020;54(2):353-364.
11. Ratitch B, Goel N, Mallinckrodt C, et al. Defining efficacy estimands in clinical trials: examples illustrating ICH E9 (R1) guidelines. *Therapeutic Innovation & Regulatory Science.* 2020;54(2):370-384.
12. Hernán MA. Does water kill? A call for less casual causal inferences. *Annals of Epidemiology.* 2016;26(10):674-680.
13. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity.* 2008;32(3):S8-S14.
14. Hernán MA, Robins JM. *Causal inference: what if.* In: Boca Raton: Chapman & Hall/CRC; 2020.
15. Hernán MA. Invited commentary: hypothetical interventions to define causal effects—afterthought or prerequisite? *American Journal of Epidemiology.* 2005;162(7):618-620.
16. VanderWeele TJ. On well-defined hypothetical interventions in the potential outcomes framework. *Epidemiology (Cambridge, Mass).* 2018;29(4):e24.
17. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology.* 2009;20(6):880-883.
18. VanderWeele TJ, Hernán MA. Causal inference under multiple versions of treatment. *Journal of Causal Inference.* 2013;1(1):1-20.
19. VanderWeele TJ. Invited commentary: counterfactuals in social epidemiology—thinking outside of "the box". *American Journal of Epidemiology.* 2020;189(3):175-178.
20. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology.* 2016;183(8):758-764.
21. Gruppen LD, Frohna AZ. Clinical reasoning. In: *International handbook of research in medical education*. Springer; 2002:205-230.
22. Groopman J. *How doctors think*. Houghton Mifflin Harcourt; 2008.
23. Rose S, van der Laan MJ. Research Questions in Data Science. In: *Targeted Learning in Data Science*. Springer; 2018:3-14.

24. McElreath R. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC; 2018.
25. Sadler DR. Formative assessment and the design of instructional systems. *Instructional science*. 1989;18(2):119-144.
26. Vandenbroucke JP. Strega, Strobe, Stard, Squire, Moose, Prisma, Gnosis, Trend, Orion, Coreq, Quorom, Remark… and Consort: for whom does the guideline toll? *Journal of Clinical Epidemiology*. 2009;62(6):594-596.
27. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal*. 2008;337.
28. Hernán MA. Counterpoint: epidemiology to guide decision-making: moving away from practice-free research. *American Journal of Epidemiology*. 2015;182(10):834-839.
29. Whittle R, Royle K-L, Jordan KP, Riley RD, Mallen CD, Peat G. Prognosis research ideally should measure time-varying predictors at their intended moment of use. *Diagnostic and Prognostic Research*. 2017;1(1):1-9.
30. Sperrin M, Martin GP, Pate A, Van Staa T, Peek N, Buchan I. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Statistics in Medicine*. 2018;37(28):4142-4154.
31. van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *European Journal of Epidemiology*. 2020;35:619-630.
32. Pajouheshnia R. *Prognostic research in treated populations*, Utrecht University; 2018.
33. Reichenbach H. The Theory of Probability: An Inquiry Into the Logical and Mathematical Foundations of the Calculus of Probability. 1949.
34. Hernán MÁ, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000:561-570.
35. Westreich D. *Epidemiology by design: a causal approach to the health sciences*. Oxford University Press; 2019.
36. Lash TL, VanderWeele TJ, Haneause S, Rothman K. *Modern epidemiology*. Lippincott Williams & Wilkins; 2020.
37. Riley RD, van der Windt D, Croft P, Moons KG. *Prognosis research in healthcare: concepts, methods, and impact*. Oxford University Press; 2019.
38. Van der Laan MJ, Rose S. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media; 2011.
39. Pfannkuch M, Forbes S, Harraway J, Budgett S, Wild CJ. *"Bootstrapping" Students' Understanding of Statistical Inference*. Teaching & Learning Research Initiative Nāu i Whatu Te Kākahu, He Tāniko Taku; 2013.
40. Morgan KL, Lock RH, Lock PF, Lock EF, Lock DF. StatKey: Online tools for bootstrap intervals and randomization tests. Paper presented at: Sustainability in statistics education. Proceedings of the 9th International Conference on Teaching Statistics, ICOTS92014.
41. Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, Reid N. Ten simple rules for effective statistical practice. *PLoS Computational Biology*. 2016;12(6):e1004961.
42. Shmueli G. To explain or to predict? *Statistical Science*. 2010;25(3):289-310.





Samenvatting in het Nederlands

In de afgelopen decennia, met name in de laatste jaren, zijn epidemiologische methoden steeds verder verfijnd, waardoor het een uitdaging is om op de hoogte te blijven van alle methodologische ontwikkelingen. De keuze voor een data-analysemethode beïnvloedt de interpretatie en klinische betekenis van resultaten direct en het is ongewenst dat het onderwerp van een onderzoek wordt bepaald door technische overwegingen. Een beter begrip van de invloed die beslissingen ten aanzien van de data-analyse kunnen hebben op de interpretatie van numerieke onderzoeksresultaten kan helpen om te komen tot de inzet van analytische instrumentaria die zowel passend als valide zijn om klinische vragen te beantwoorden. Het doel van dit proefschrift is om de invloed van keuzes omtrent het ontwerp en de statistische analyse van een studie op de betekenis van de numerieke resultaten te onderzoeken, met specifiek aandacht voor onderzoek naar causale effecten (Deel I) en voorspellingsonderzoek (Deel II).

Deel I: Invloed van toegepaste methoden op de betekenis van numerieke schattingen in onderzoek naar causale effecten

De studies beschreven in Deel I van dit proefschrift zijn voorbeelden van de invloed van keuzes omtrent het studie ontwerp en de statistische analyse op numerieke resultaten in onderzoeken naar causale effecten. Dit deel schetst de vele data-analytische beslissingen die in zulke onderzoeken dienen te worden genomen. In de hoofdstukken 2 en 4 worden twee specifieke beslissingen uitgelicht en is de conclusie dat de interpretatie van effectschattingen in onderzoeken naar causale effecten in belangrijke mate afhangt van de keuze voor respectievelijk de studietijdoorsprong en de strategie ten aanzien van selectie van covariabelen. De resultaten van deze onderzoeken impliceren dat het belangrijk is om een tegengesteld proces te vermijden, waarbij de toegepaste onderzoeksopzet en statistische analyse impliciet de betekenis van de onderzoeksresultaten bepalen. In plaats daarvan zou juist eerst het doel van het onderzoek duidelijk moeten zijn, worden vastgesteld met welke onderzoeksresultaten die doelen kunnen worden bereikt om pas daarna te bepalen welke data-analyse daarvoor geschikt is. Dit wordt benadrukt door de onderzoeken beschreven in de hoofdstukken 3 en 5, waarin wordt besproken hoe een gericht geformuleerd onderzoeksdoel kan verduidelijken of de klinische doelen, de onderzoeksopzet en de

interpretatie van resultaten op de juiste manier op elkaar zijn afgestemd. Het vaststellen van het onderzoeksdoel is echter een uitdaging in klinische onderzoeken. In hoofdstuk 2 wordt deze uitdaging weerspiegeld in de lage frequentie van expliciet gerapporteerde *estimands*, ofwel de 'te schatten waarde die de klinische onderzoeksvraag waarin men geïnteresseerd is beantwoordt (of het beste benadert)', in onderzoeken naar effecten van farmacologische behandelingen.

Hoofdstuk 2 beschrijft dat de invloed van de gekozen studietijdsoorsprong op numerieke resultaten moeilijk te beoordelen was in farmaco-epidemiologische studies vanwege onvolledige rapportage. De rapportage over keuzes met betrekking tot het operationaliseren van de tijdsoorsprong van een studieontwerp werd onderzocht in een systematische review van 89 vergelijkende cohortstudies voor effectiviteit en veiligheid, gepubliceerd in zes hoog aangeschreven farmaco-epidemiologische tijdschriften in 2018 en 2019. Veertig procent van de onderzoeken rapporteerde de studie inclusie te beperken tot alleen nieuwe gebruikers (d.w.z., een 'incidentgebruikersontwerp') en 13% rapporteerde ook bestaande gebruikers te includeren (d.w.z., een 'prevalentgebruikersontwerp'). De start van opvolging binnen de studie, het moment waarop aan de toelatingscriteria werd voldaan en de start van de behandeling bleken overeen te komen in 53% van de onderzoeken met een incidentgebruikersontwerp (19 van de 36 onderzoeken) en werd in 42% van de onderzoeken met een incidentgebruikersontwerp onvoldoende beschreven. De validiteit van de operationalisering van de studietijdsoorsprong kon alleen worden beoordeeld met betrekking tot de *estimand*. De *estimand* werd echter in slechts 22% van de onderzoeken met een incidentgebruikersontwerp expliciet gerapporteerd.

In **Hoofdstuk 3** wordt de nauwkeurigheid waarmee het onderzoeksdoel wordt gedefinieerd in explorerende etiologische studies gekoppeld aan de interpretatie van de bevindingen van dergelijke studies. Er wordt een continuüm beschreven van kritische evaluatie voor het uitvoeren van studies, variërend van *ad-hoc* tot *doordacht*. De positie waarop een explorerende etiologische analyse zich op dit continuüm bevindt, beïnvloedt direct de interpreteerbaarheid van bevindingen. Er wordt beargumenteerd dat het handelen naar resultaten van *ad-hoc* analyses alsof ze voortkomen uit *doordachte* analyses, bijvoorbeeld door verdere bevestigende studies uit te voeren of door ze in de klinische praktijk te implementeren, kan bijdragen aan verspilling van onderzoeksmiddelen en patiënten mogelijk kan schaden. Er worden praktische

aanwijzingen gegeven voor een goede uitvoering van explorerend etiologisch onderzoek, zoals het gebruik van rigoureuze methodologische en statistische benaderingen en het nemen van verantwoordelijkheid voor exploratieve bevindingen door een duidelijke agenda voor toekomstig onderzoek te rapporteren.

De studie beschreven in **Hoofdstuk 4** illustreert dat het toepassen van achterwaartse eliminatie om het aantal covariabelen waarmee voor vertroebeling (*confounding*) gecorrigeerd wordt te verminderen zelden efficiënter is dan selectie van covariabelen op basis van causale kennis. Er wordt een uitdrukking afgeleid die kwantificeert hoe het weglaten van een variabele zich verhoudt tot vertekening (*bias*) en variantie van effectschatters. Er zijn simulaties uitgevoerd om te onderzoeken of en onder welke omstandigheden de schatting van causale effecten in observationele onderzoeken kan worden verbeterd door achterwaartse eliminatie te gebruiken op een vooraf gespecificeerde set van potentiële *confounders*. Het toepassen van achterwaartse eliminatie verminderde de gemiddelde fout van effectschatters niet in vergelijking met een volledig model inclusief alle vooraf gespecificeerde covariabelen, maar de *bias* was toegenomen. In minder dan 3% van de 3960 beschouwde scenario's was de gemiddelde fout van effectschatters lager wanneer achterwaartse eliminatie werd gebruikt in vergelijking met het volledige model. Deze bevindingen geven daarmee aan dat wanneer een eerste set van potentiële *confounders* kan worden gespecificeerd op basis van achtergrondkennis, er een minimale toegevoegde waarde is van achterwaartse eliminatie.

In **Hoofdstuk 5** wordt een methode voorgesteld om studies die in systematische reviews van chirurgische interventies worden geïncludeerd te beoordelen met betrekking tot keuzes omtrent het ontwerp en de statistische analyse. Dit is bedoeld als een eerste stap richting het samenvatten van de belangrijkste informatie die nodig is om de toepasbaarheid en methodologische kwaliteit van studies te beoordelen. Daartoe is – op basis van bestaande instrumenten die het risico op bias beoordelen – een beknopte set van items gekozen voor de initiële evaluatie van studies van chirurgische interventies. De set bevat negen items: populatie, interventie, comparator, uitkomst, *confounding*, ontbrekende gegevens en selectiebias, interventiestatus, uitkomstbeoordeling en pre-specificatie van de analyse. De beoordeling van toepasbaarheid en methodologische kwaliteit kan worden gedaan als onderdeel van een systematische review om een eerste schifting te maken waarin studies van lage kwaliteit relatief gemakkelijk kunnen worden uitgesloten. Studies van hogere kwaliteit kunnen aan verdere beoordeling

van methodologische kwaliteit worden onderworpen met behulp van bestaande beoordelingsinstrumenten.

Deel II: Invloed van toegepaste methoden op de betekenis van numerieke schattingen in voorspellingsonderzoek

De studies beschreven in Deel II van dit proefschrift zijn gericht op voorspellingsonderzoek en bestuderen de invloed van veranderingen in strategieën voor het meten van voorspellers tussen verschillende modelleringsfasen op de mate waarin voorspelmodellen correcte voorspellingen doen (d.w.z., de 'modelprestaties'). Dergelijke veranderingen worden *voorspeller-meetheterogeniteit* genoemd. Het fenomeen *voorspeller-meetheterogeniteit* werd in dit proefschrift formeel gedefinieerd met behulp van meetfoutmodellen. Hierdoor was het mogelijk de implicaties van *voorspeller-meetheterogeniteit* te onderzoeken door middel van analytische benaderingen, simulaties, analyse van empirische datasets en een voorgestelde kwantitatieve voorspellingsanalyse. Al deze analyses geven aan dat zelfs wanneer alle andere factoren, zoals de modelleringsstrategie, uitkomstprevalentie, voorspellers en patiëntkenmerken, constant worden gehouden in alle modelleringsfasen, een verandering in de meetprocedure de modelprestaties beïnvloedt. Dit nodigt uit tot een heroverweging van de manier waarop voorspellingsmodellen worden gespecificeerd, en met name of voorspellermeetprocedures een integraal onderdeel van de modelspecificatie zouden moeten zijn.

Hoofdstuk 6 beschrijft hoe strategieën voor het meten van voorspellers relateren aan klinische toepasbaarheid van voorspellingen van binaire logistische voorspellingsmodellen met behulp van analytische- en simulatiebenaderingen. Een gevestigde taxonomie van meetfoutmodellen wordt gebruikt om het fenomeen *voorspeller-meetheterogeniteit* te definiëren en te verduidelijken. Voorspeller-meetheterogeniteit refereert naar variatie in de meetprocedure van voorspeller(s) tussen de data waarin een voorspellingsmodel wordt afgeleid en waarin een model wordt gevalideerd of geïmplementeerd. Met behulp van analytische en simulatiebenaderingen wordt aangetoond dat modelprestaties buiten de steekproef van binaire logistische voorspellingsmodellen kunnen worden belemmerd wanneer voorspellers anders worden gemeten bij afleiding en externe validatie. Deze bevindingen benadrukken dat het onvoldoende is om een voorspellingsdoel in algemene termen te beschrijven zonder de procedures te specificeren waarmee voorspellers (moeten) worden gemeten.

In **Hoofdstuk 7** wordt getoond hoe strategieën voor het meten van voorspellers relateren aan klinische toepasbaarheid van voorspellingen van binaire logistische diagnostische modellen met behulp van empirische illustraties in drie klinische datasets. Negen scenario's van *voorspeller-meetheterogeniteit* worden geëvalueerd in eerder ontwikkelde voorspellings-modellen voor de diagnose van eierstokkanker, mutatiedragers voor Lynch-syndroom en intra-uteriene zwangerschap. Het wijzigen van de meetprocedure van een voorspeller had invloed op de modelprestaties bij validatie van de diagnostische modellen, met name modelkalibratie, waarbij de coëfficiënt voor gemiddelde kalibratie bij validatie varieerde van -0,70 tot 1,43 en de kalibratiehelling van 0,50 tot 1,67.

Hoofdstuk 8 beschrijft een kwantitatieve voorspellingsanalyse om te anticiperen op de invloed van veranderingen in voorspellermeetstrategieën voor prognostische tijd-tot-gebeurtenis uitkomstmodellen. Met behulp van simulaties met verschillende scenario's van *voorspeller-meetheterogeniteit*, hebben we aangetoond dat modelprestaties buiten de steekproef kunnen worden belemmerd wanneer voorspellers anders worden gemeten bij validatie en implementatie voor tijd-tot-gebeurtenis uitkomstmodellen. Er wordt een kwantitatieve voorspellingsanalyse voorgesteld om de invloed te kwantificeren van de verwachte *voorspeller-meetheterogeniteit* over de validatie- en implementatiesetting heen.

Tot slot

Dit proefschrift beschrijft de invloed van verschillende keuzes omtrent het ontwerp en de statistische analyse van een studie op de betekenis van de numerieke resultaten. Het duidelijk definiëren van een klinisch relevante *estimand* zorgt ervoor dat data-analytische beslissingen zinvolle resultaten opleveren. Echter, het vereist een diepgaand begrip van statistische theorie om een *estimand* te definiëren en hiervoor zijn weinig aanbevelingen beschikbaar, met name niet voor niet-therapeutisch onderzoek.

In **Hoofdstuk 9**, de algemene discussie van dit proefschrift, stellen we voor hoe dergelijke aanwijzingen zouden kunnen worden ontwikkeld om onderzoekers te helpen tot een voldoende goed gedefinieerde *estimand* te komen, te beginnen vanuit een klinisch perspectief, dat wil zeggen door het definiëren van een *´doordachte onderzoeksvraag'*. Vijf onderwerpen voor toekomstig onderzoek worden voorgesteld die kunnen bijdragen aan het centraal stellen van *doordachte onderzoeksvragen* in klinisch onderzoek, te weten:

(1) Onderscheid een klinisch perspectief van een statistisch perspectief met als doel de input van beide op elkaar af te stemmen.
(2) Identificeer de optimale balans tussen beknoptheid en volledigheid van *doordachte onderzoeksvragen*.
(3) Specificeer hoe en wanneer *doordachte onderzoeksvragen* verschillen tussen onderzoeken naar causale effecten en voorspellingsonderzoek.
(4) Voer empirische voorbeeld-studies uit om te begrijpen hoe gerichte onderzoeksvragen gesteld kunnen worden.
(5) Geef les in het stellen van *doordachte onderzoeksvragen*.

Door *doordachte onderzoeksvragen* centraal te stellen in kwantitatief klinisch onderzoek kan voorbarig vertrouwen in (complexe) methoden worden verminderd en kan het begrip van de waarde en betekenis van onderzoeksresultaten worden vergroot.

PhD portfolio

Oral presentations

New-user and prevalent-user designs and the definition of study time origin in pharmacoepidemiology: a review of reporting practices — 2021
International Society for Clinical Biostatistics, Lyon, France (online)

New-user and prevalent-user designs and the definition of study time origin in pharmacoepidemiology: a review of reporting practices — 2021
Annua Dutch Epidemiology Conference, WEON, Amsterdam, the Netherlands (online)

Automated covariate selection for causal inference? — 2019
STRATOS meeting at Institut für Medizinische Biometrie, Informatik und Epidemiology, Bonn, Germany

Measurement matters: how differences in measurement induce non-transportability of clinical prediction models — 2019
International Society for Clinical Biostatistics, Leuven, Belgium

Impact of predictor measurement heterogeneity on performance of clinical prediction models: a measurement error perspective — 2018
Methods for Evaluation of medical prediction Models, Tests and Biomarkers, Utrecht, the Netherlands

Invited seminars

Incident and prevalent-user designs and the definition of study time origin in pharmacoepidemiology — 2021
Seminar at Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht, the Netherlands (online)

Measurement matters: how differences in measurement induce non-transportability of clinical prediction models — 2019
Seminar at Center for Medical Statistics, Informatics, and Intelligent Systems, Vienna, Austria

Am I ready to generate scientific output? Finding a balance in learning *by* doing and learning *before* doing — 2019
International Society for Clinical Biostatistics Early Career Biostatistician's Day, Leuven, Belgium

Poster presentations

Selection of covariates for confounding adjustment in studies of causal inference: backward elimination of covariates had minimal added value to using background knowledge 2020
International Society for Clinical Biostatistics, Krakow, Poland (online)

Measurement matters: how differences in measurement induce non-transportability of clinical prediction models 2019
Annual Dutch Epidemiology Conference, WEON, Groningen, the Netherlands

Publications

Luijken K, Song J, Groenwold RHH, Quantitative prediction error analysis to investigate predictive performance under predictor measurement heterogeneity at model implementation. *Diagnostic and Prognostic Research* (in press).

Luijken K, Groenwold RHH, van Smeden M, Strohmaier S, Heinze G. A comparison of full model specification and backward elimination of potential confounders when estimating marginal and conditional causal effects on binary outcomes from observational data. *Biometrical Journal* (in press).

Luijken K, Dekkers OM, Rosendaal FR, Groenwold RHH. Exploratory analyses in etiologic research: considerations for assessment of credibility. *BMJ* (in press).

Luijken K, Spekreijse JJ, van Smeden M, Gardarsdottir H, Groenwold RHH. New-user and prevalent-user designs and the definition of study time origin in pharmacoepidemiology: A review of reporting practices. *Pharmacoepidemiology and Drug Safety*. 2021;30(7):960-74.

Hempenius M, Luijken K, de Boer A, Klungel O, Groenwold RHH, Gardarsdottir H. Quality of reporting of drug exposure in pharmacoepidemiological studies. *Pharmacoepidemiology and Drug Safety*. 2020;29(9):1141-50.

van Geloven N, Swanson SA, Ramspek CL, Luijken K, van Diepen M, Morris TP, Groenwold RHH, van Houwelingen HC, Putter H, le Cessie S. Prediction meets causal inference: the role of treatment in clinical prediction models. *European Journal of Epidemiology*. 2020;35:619-30.

Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MM, Dahly DL, Damen JA, Debray TP, …, van Smeden M. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. 2020;369:m1328.

Luijken K, Wynants L, van Smeden M, Van Calster B, Steyerberg EW, Groenwold RHH. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *Journal of Clinical Epidemiology*. 2020;119:7-18.

le Cessie S, Luijken K, Goetghebeur E. Regarding "Variable selection-A review and recommendations for the practicing statistician" by G. Heinze, C. Wallisch, and D. Dunkler". *Biometrical Journal*. 2019;61(6):1595-1597.

Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. *Statistics in Medicine*. 2019;38(18):3444-59.

Paauw ND, Luijken K, Franx A, Verhaar MC, Lely AT. Long-term renal and cardiovascular risk after preeclampsia: towards screening and prevention. *Clinical Science*. 2016;130(4):239-46.

Curriculum Vitae

Kim Luijken was born in Rotterdam, the Netherlands, on the 26th of April 1994. She graduated from the Regius College in Schagen in 2012. After completing a bachelor's degree in medicine (with distinction) at Utrecht University, she switched to studying methodology and statistics at the same university. She worked on her master's thesis at the Department of Clinical Epidemiology at the Leiden University Medical Center which was awarded the Statistics Netherlands Award 2018. In 2018, she graduated as a Master of Science in applied methodology and statistics (with distinction) and started a PhD under the supervision of Prof. dr. R.H.H. Groenwold. By participating in (inter)-national courses, journal clubs, and conferences, she broadened her understanding of studies of causal inference and prediction modelling studies. Her methodological work was provided impetus by a research visit to the University of Vienna and a collaboration with the Catholic University of Leuven. At the time of writing, Kim is working as a researcher at the Department of Epidemiology at the Julius Center, University Medical Center Utrecht.

Dankwoord

Dit proefschrift is een weerspiegeling van alles wat ik gedurende mijn PhD traject heb geleerd, zowel in de methodologie als op persoonlijk vlak. Graag bedank ik iedereen die aan deze ontwikkeling heeft bijgedragen.

Rolf, dank voor je energieke, scherpzinnige, en motiverende begeleiding. Toen ik vier jaar geleden mijn twijfels met je deelde of ik met deze PhD wilde beginnen, zette je die met een relativerende grap om in motivatie om te starten. De toon was gezet. Zo begon een traject waarin je me veel ruimte gaf projecten naar eigen inzicht vorm te geven en tegelijkertijd betrokken begeleiding bood. Bedankt voor de aandacht die je in deze begeleiding gestoken hebt en het verrijkende traject waarin dat resulteerde!

Maarten, dankjewel voor je leerzame en hartelijke begeleiding en voor je zorgzaamheid op momenten waarop ik belangrijke keuzes moest maken. Je doordachte commentaren maakten juist die extra slag in een manuscript.

Linda en Bas, het was enorm fijn om deze PhD samen met jullie te doorlopen en om inzichten, strubbelingen en gezelligheid te delen. Ik heb veel van jullie geleerd en daarbij volop plezier gehad. Bedankt!

Laure, Ben en Ewout, de projecten die uit onze samenwerking zijn voortgekomen hebben een vliegende start gegeven aan mijn PhD! Bedankt voor jullie interesse in predictor meetheterogeniteit, jullie scherpe inzichten en de gezellige besprekingen. Georg, and everyone at the CeMSIIS group, thank you for an intriguing and educational project in computational epidemiology – and for the Sturm on hikes through Viennese surroundings! Marijn en Bryan, in het laatste jaar van mijn PhD deed zich een samenwerking voor met een voor mij verrassende richting: de traumatologie! Deze onderzoeken brengen me verbreding, verdieping en plezier; bedankt daarvoor.

I am grateful for the inspiring and gezellige time with everyone at the Department of Clinical Epidemiology at the LUMC. Thank you! Tamara en Yvonne, dank voor al jullie organisatorische inzet.

The realization of this dissertation would not have been as smooth and a lot less fun without the nifty tools I picked up during our Hacky Hours! Thanks Anna, Daniela, Ed, Linda, Tariq, and Xante.

Lieve Judith, ik ben je dankbaar dat je met jouw expertise als betrokken en doortastende arts aan hoofdstuk 2 van dit proefschrift hebt bijgedragen. En voor de wandelingen, zwemuurtjes en diners!

Dear class of 2018, team Be(e)rnoulli, especially Liz, Felix, Jolien, Nicolas, Dario, Jeroen, Lieke, Anna-Sophia, Elian, and Juan. Thank you for making statistics even more fun with trips to the SGG, Geneva, Berlin, the Bata, Mexico, and Colombia. I look forward to further exploring the world of methodology and statistics together with you guys – as well as to completely forget about it every now and then.

Het is fijn om af en toe helemaal uit dit werk te stappen en terug te gaan naar de roots. GZA, lieverds, bedankt voor jullie vrolijkheid en de energie die jullie me geven.

Lieve Richard, bedankt voor alle inspirerende en warmhartige gesprekken om het leven lichter en feestelijker te maken. Lieve Timna, zo verschillend als de wegen zijn die we al deze jaren bewandelen, zo vreugdevol vind ik de raakvlakken die steeds weer opduiken. Jullie beider vriendschap is me erg dierbaar. Wat fijn dat jullie vandaag als paranimfen naast me staan!

Lieve pap en mam, bedankt voor jullie warmte en bemoediging. Pap, jij wist al voordat ik hieraan begon wat het onderwerp van mijn PhD zou worden. Ik had er toen nog geen oren naar, maar je liet het me rustig zelf uitvogelen. Mama, toen ik wilde switchen van geneeskunde naar methodologie steunde jij me meteen vol enthousiasme. Dat heeft me de moed gegeven de overstap te maken en daar heb ik zoveel plezier van!

Liefste Rick, jouw onvoorwaardelijke begrip, vrolijke humor en vertrouwen zijn een enorm geschenk. Ik geniet elke dag van ons.