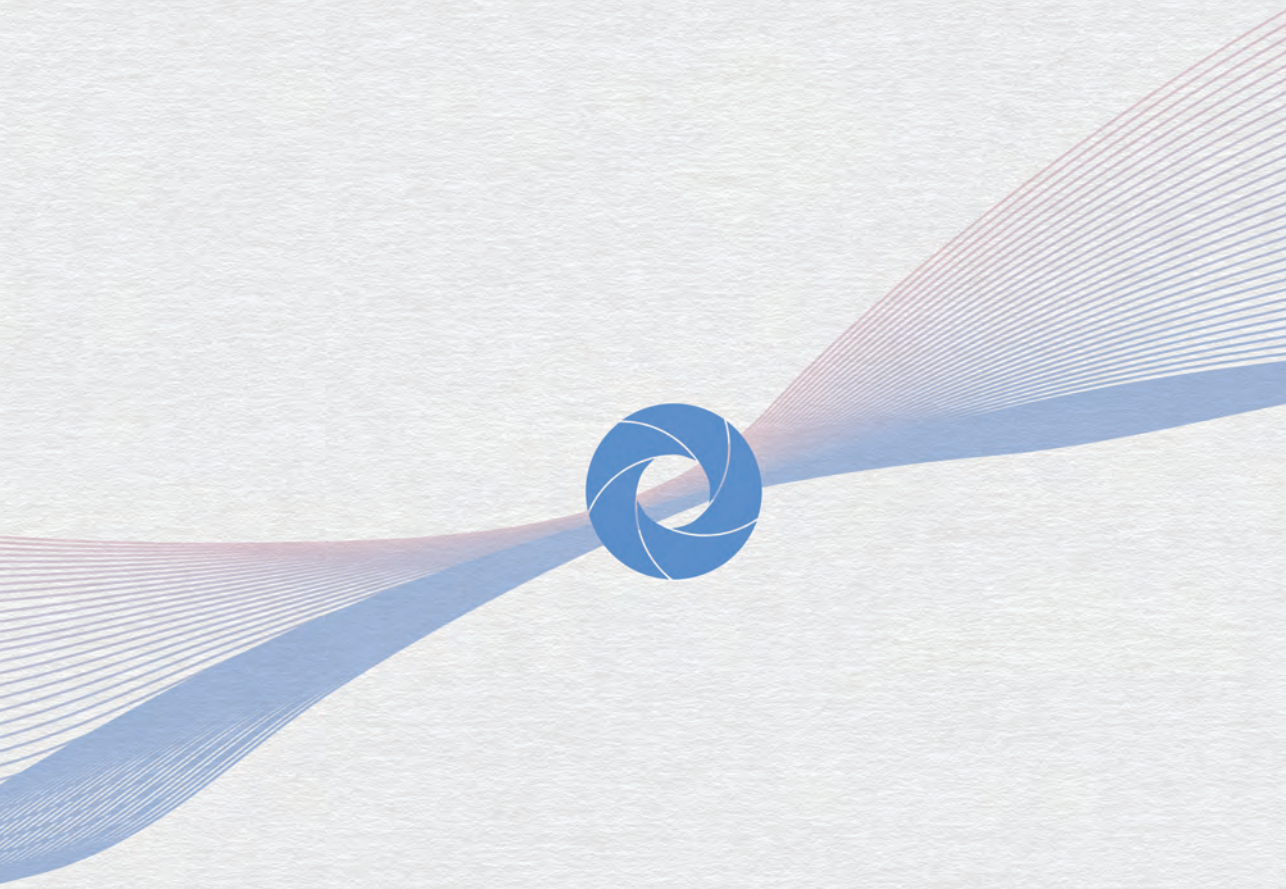
The impact of epidemiologic methods on findings in studies of causal effects and prediction modelling

Luijken, K.

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].

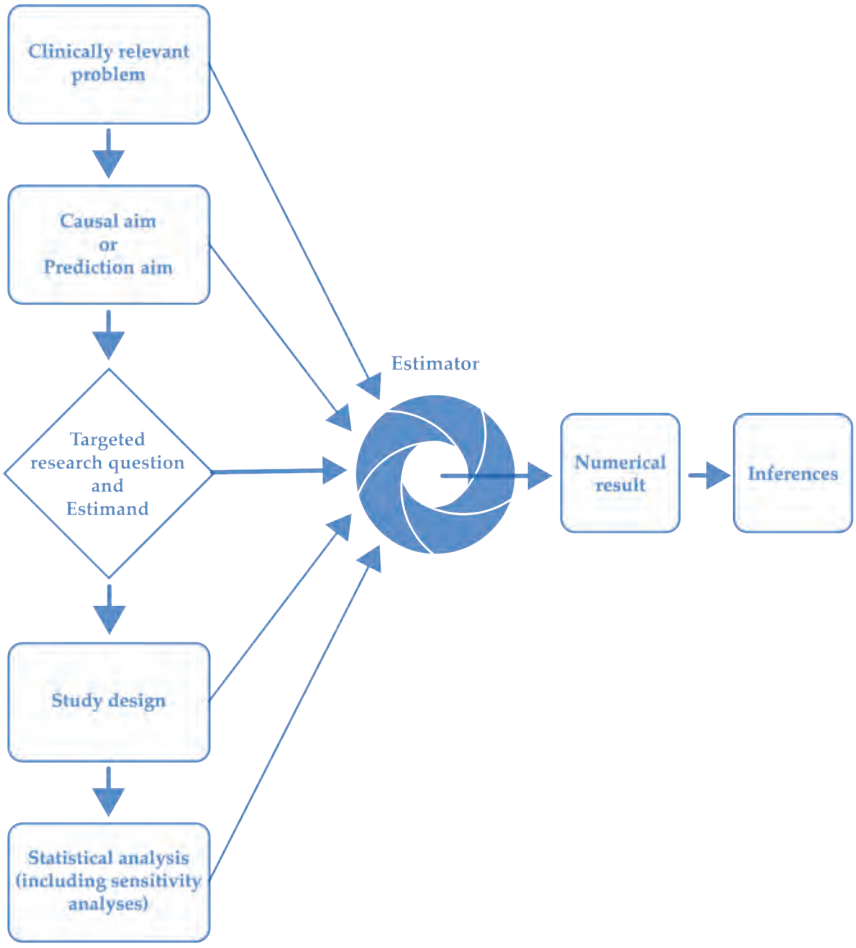


Figure 1. Adding a missing link to the schematic depiction of stages of research conduct of a quantitative study. When data analytical decisions are not driven by substantial clinical considerations, this could result in a disparity between the clinical research question and the meaning of a numerical 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.