Methodological obstacles in causal inference: confounding, missing data, and measurement error

Penning de Vries, B.B.L.

12

Summary and general discussion

Epidemiology is a broad field of study with methods and concepts connecting all subfields (Lau et al., 2020). This thesis describes a study of epidemiological methods for answering questions about cause and effect in the presence of methodological obstacles, such as confounding, missing data or measurement error. In this chapter, a summary of our main findings is presented, along with a general discussion of this thesis in the light of the existing literature, with suggestions for future research.

12.1 Summary of findings

Methods for answering causal questions can be studied with the aim of learning about its workings, its performance under certain conditions. At a more meta-level, we can study how methods are being disseminated or implemented. Likewise, we can study, on the one hand, how and when a methodological obstacle may be overcome, and, on the other, how it is handled in applied research. In **chapter 2**, we questioned some of the current practice of how research at this meta-level is conducted, particularly where it concerns the initial phases of a systematic literature review. The standard approach of ignoring the text body in searching or screening articles might fail to retrieve all or a representative sample of the relevant literature, potentially leading to a false impression about the topic of enquiry. We found that for a number of methodological topics, a large portion of articles with a topic mention somewhere in the text did not contain a reference to the topic in text fields other than the body. The results do not conclusively show that ignoring text bodies does indeed lead to a false impression, but it

should raise suspicion. Researchers might wish to consider including these text fields in their search and screening strategy.

In primary research, epidemiologists are often faced with multiple methodological obstacles simultaneously. There are concerns, however, that combinations of methods designed for different methodological obstacles have worse performance than might be expected from how they perform in isolation. In **chapters 3 and 4**, we critically reflected on a previous simulation study by Mitra and Reiter (2016), in which they compare two approaches to implementing propensity score matching after multiply imputing missing data. We found that the standard multiple imputation approach of carrying out analysis *within* multiply imputed datasets before pooling the results is generally to be preferred over their proposed approach of first pooling propensity scores *across* multiply imputed datasets before carrying out matching (or any other propensity score method) based on these pooled scores. Our results are in stark contrast to the results of Mitra and Reiter (2016) and we argued that this is largely due to their use of a misspecified imputation model that ignores the outcome variable.

Propensity score estimation is typically done by fitting a logistic regression. However, standard regression modelling software by default discards all incomplete records and does not offer propensity score estimates for subjects with missing data. Machine learning techniques such as classification and regression trees (CART) are appealing in part because some implementations allow for incomplete records to be incorporated in the tree fitting and provide propensity score estimates for all subjects. An important question to be answered is whether and when CART handles the missing data in a desirable way. In **chapter 5**, we argued that the automatic handling of missing data by CART is by no means a one-fits-all solution to the problem of missing covariate data for causal inferences based on propensity score methods. In a number of simulation studies, we actually found CART to be outperformed by standard, alternative methods to account for missing data. Different CART implementations handle missing data differently. In judging whether a given implementation is appropriate for the task at hand, some understanding of the 'black-box nature' of machine learning algorithms is therefore desirable.

In **chapter 6**, we considered missing outcome rather than missing covariate data. The chapter gives no new results but emphasises and illustrates that when baseline exchangeability is achieved through propensity score matching, bias might nonetheless result from restricting downstream analysis to the subset of individuals who have not dropped out of the study by the administrative study end. This equally applies to controlled trials with baseline randomisation, where exchangeability, achieved at baseline by design, is not guaranteed to uphold in

the set of complete records that may be used for the analysis. Regression and inverse probability of censoring weighting were discussed as possible solutions.

Researchers can sometimes have a considerable influence over the extent of missingness. In studies on the effects of time-varying exposures, information of post-baseline covariates may help mitigate time-dependent confounding, but obtaining a record of the values that these variables take at each of potentially many time points can be costly and time-consuming. Reducing the frequency of measurements may enhance study feasibility, but it may also compromise study validity. In **chapter 7**, we illustrated by way of simulation the impact of choices regarding the frequency of measuring time-varying covariates. To handle missing values, we implemented the last-observation-carried-forward procedure (LOCF) under the implicit (and wrong) assumption that the participant characteristics remained constant in periods of no measurement. As expected, in our simulations, fixed-interval measurement resulted in bias consistent with residual confounding. We additionally showed that bias might arise in settings where decisions to measure are driven by observed values of the time-varying exposure, such as in the studies of Ali et al. (2016) and Souverein et al. (2016).

When variables take values that are different from what these values appear or are assumed to be, such as may be the case when we implement LOCF, we say that the variables are subject to measurement error. When the variables are categorical, we speak of misclassification, a special type of measurement error. In **chapter 8**, we focused on joint exposure-outcome misclassification and developed a method for this issue in the presence of confounding. Simulation studies showed favourable large sample performance. However, further research is needed to study the sensitivity of the proposed method and that of alternatives to violations of their assumptions.

Concerns about violations of assumptions are common in observational research on causal effects. In efforts to lessen these concerns, it has been suggested that so-called negative control variables are used (Lipsitch et al., 2010). Negative controls are variables that are known (or at least believed) to be causally unrelated to one or more of the variables of interest. The key idea is that observing an association that contradicts the belief in a causal null relation might alert the analyst to violations of assumptions. Negative controls have potential in bias detection as well as partial or complete bias correction in epidemiological research. In **chapter 9**, we sought to complement efforts to increase the more routine use of negative controls with a discussion about a selection of caveats. We argued that negative controls may lack both specificity and sensitivity to detect unmeasured confounding. We also reviewed existing methods to adjust for unmeasured confounding based on negative controls and examined the impact

of assumption violations. Given the potentially large impact, it may sometimes be desirable to replace strong conditions for exact identification with weaker, easily verifiable conditions, even when these imply at most partial identification. Future research in this area may broaden the applicability of negative controls and in turn make them better suited for routine use in epidemiological practice. At present, however, the applicability of negative controls should be carefully judged on a case-by-case basis.

Case-control designs are an important tool in causal inference. In **chapter 10**, we argued that to facilitate understanding, it is useful to consider every case-control study as being nested within a cohort study. The case-control study then effectively becomes a cohort study with missingness governed by the control-sampling scheme. In the chapter, we gave an overview of how observational data obtained with case-control designs can be used to identify a number of causal estimands and, in doing so, recast historical case-control concepts, assumptions and principles in a modern and formal framework.

Finally, in **chapter 11**, we turned to precision medicine and considered the task of finding the optimal subgroup for treatment under certain cost or resource constraints. In practice, it is not uncommon for treatment assignment decisions to be based of prognostic scores. However, this approach does not guarantee optimal results (VanderWeele et al., 2019). As an alternative, one may attempt to evaluate all possible subgroups one by one, and choose the rule with the 'best' results. However, this is not feasible when there are many, potentially infinitely many subgroups to consider. VanderWeele et al. (2019) showed that the task can sometimes be considerably simplified by deriving treatment assignment rules that (1) guarantee optimality under some conditions and (2) take a simple form: assign treatment in a greedy fashion to all individuals with the next largest benefit (i.e., the largest difference in potential outcome means given covariates) or the next highest benefit–cost ratio (with cost being a positive function of baseline covariates) until the resource or cost constraint, respectively, is exceeded. The optimality of the rules however relies critically on the assumption that there are no tied conditional treatment effects or benefit-cost ratios between individuals. We extended their work by deriving rules that likewise have a simple form and which guarantee optimality under the same conditions, except that there need be no constraint on the presence of ties. An important insight that this chapter is meant provide is that in order to obtain some sense of optimally from allocating treatment, a contrast between counterfactual outcomes under different treatment options should be considered. Prognostic scores alone are not (generally) sufficient. The methodological obstacles that we encounter in causal inference, including confounding, missing data and measurement error,

are therefore relevant in precision medicine too.

12.2 General discussion

The methodological aspects of causal inference form a broad topic and we addressed a variety of subtopics in this thesis. Apart from confounding, missing data, and measurement error, the reader may nonetheless recognise a number of recurrent features.

For example, Monte Carlo simulation was used in a number of chapters (e.g., **chapters 3-8**). This is a useful tool for obtaining empirical results (i.e., approximations) about the performance of statistical methods in certain scenarios as opposed to more general, analytic results (Morris et al., 2019). They are particularly appealing when the latter are difficult to obtain, or when the interest lies with illustrating a problem or method. However, they also have limitations. They provide at most approximations of statistical properties. Also, only a limited, finite number of scenarios can be considered and there is often the concern that the results generalise poorly to other scenarios.

Much of this thesis is built on the potential or counterfactual outcomes framework. In this work, like much of the literature, the terms 'potential outcomes' and 'counterfactual outcomes' are used interchangeably. Where they are considered distinct, generally the potential and counterfactual versions of a variable under the same hypothetical situation are still regarded as having the same values. However, variables are labeled as potential or counterfactual depending on whether they are seen as primitive or constructed from a collection of functions and background variables, respectively (Pearl, 2010). In some parts of this thesis (e.g. **chapter 5**), we explicitly took a constructivist approach, while in others (e.g., **chapter 10**), we did not. The adjective 'potential' further connotes a prospective view; either one of multiple versions of the outcome might become real-world before the choice among the corresponding mutually exclusive actions is made. By contrast, 'counterfactual' connotes a retrospective view; the choice among mutually exclusive actions is made and all but one version of the outcome is contrary-to-fact.

The notion of 'counterfactual thinking' is not used merely in epidemiology and has found its way in many branches of science, including physics (Robins et al., 2015). Its uptake and popularity in epidemiology, however, have given rise to much dispute among academics (Vandenbroucke et al., 2016; Krieger and Davey Smith, 2016; Broadbent et al., 2016; VanderWeele, 2016; VanderWeele et al., 2016; Schwartz et al., 2016; Daniel et al., 2016; Robins and Weissman, 2016;

Blakely et al., 2016). A central point of critique is that counterfactual thinking would delimit the meaning of causality by equating "causal claims with precise predictions about contrary-to-fact scenarios" (Vandenbroucke et al., 2016). A contrasting view is that the counterfactual framework considers a subset—not necessarily the entire set—of causal claims, namely those that can be phrased as statements about the consequences of hypothetical—possibly contrary-to-fact—actions (VanderWeele et al., 2016). Sometimes, the framework might admit non-actions (e.g., states) as potential causes but only when it is understood what actions are implied. The focus on this subset of causal claims is meant to guide decisions in the real world based on predictions of their consequences.

It should be noted that even after restricting to this subset of causal questions, some ambiguity about what the actions (interventions) and corresponding counterfactuals mean often remains. This issue relates to another point of debate: the well-definedness of interventions and counterfactuals. It is important to note that well-definedness of interventions is not the same as the interventions being elaborate. Telling a patient to follow a poorly detailed drug prescription or exercise programme, and advising social distancing against the spread of COVID-19 during a given press conference may well represent reasonably well-defined (point) interventions. They are not made less well-defined by the patient being unsure of how to interpret the drug prescription or exercise programme, or by the residents of a country not acting on the social distancing advise in a uniform way. Well-definedness of interventions relates to the lack of ambiguity of what the interventions mean, not about how they should be acted on. The requirement that interventions and counterfactuals are sufficiently well-defined, as noted in the introduction of this thesis, is that there is no ambiguity about the interventions or that the counterfactuals are invariant to the choice among the possible variations. Striving for well-definedness only serves to eliminate vagueness about the meaning of a causal effect.

Other critique relates to the assumptions that can be readily made explicit with counterfactual parlance (Schwartz et al., 2016), and the tools that are typically associated with or embedded in the counterfactual framework, such as directed acyclic graphs (DAGs) or single-world intervention graphs (sWIGs) (Richardson and Robins, 2013) with which some assumptions can be graphically encoded. However, that the assumption of, say, 'no interference' for a joint intervention on multiple individuals (i.e., 'one individual's treatment does not affect another's outcome') is often made (albeit often implicitly) or can be articulated with relative ease, does not mean that the counterfactual framework permits only causal inference under this assumption (Robins and Weissman, 2016). The development of a language rich enough to articulate a wider variety

of causal questions and assumptions is an advance with positive effects on clarity of thought and ease of communication. The assumptions that are made explicit and least ambiguously articulated are inevitably often the ones that receive the most scrutiny and criticism. As Pearl et al. (2014) notes, "he who seeks licensing assumptions risks suspicions of attempting to endorse those assumptions. ... The more explicit the assumption, the more criticism it invites". Methodological decisions (e.g., about which variables to 'adjust' for, or about the use of complete case analysis versus multiple imputation for missing data) often rely on structural assumptions about the data. There are often concerns that the DAGs encoding data structures are too simplistic. Robins (1999) argues that although the real world may well be more complex than is sometimes implied by a simple graph, "if we do not learn how to reason correctly in simple causal Gedankenexperiments ..., we have no chance of success in realistic situations." Uncertainty about whether certain (identifiability) assumptions are met does not justify that potential assumption violations are ignored or rigour abandoned.

Like 'counterfactuals', 'missing data' and 'measurement error' are terms whose meaning is not always clear. For example, it is easy to conflate a given variable being inaccessible to the researcher (often encoded with 'NA') with the variable being accessible yet taking the value 'missing' or 'NA'. For example, in an attempt to address confounding, one might wish to capture all information upon which a general practitioner (GP) bases his treatment decisions. The GP might fail to take a patient's blood pressure, but this does not mean that the corresponding variable is truly missing. The GP cannot base decisions on what he did not observe and, so, the researcher might still have access to all variables that have informed the GP's decision making. Similar comments apply to the notion of measurement error. Measurement error is a relative notion: in one context, systolic blood pressure plus some random term might be considered measurement error; in others, it is exactly what the researcher set out to measure.

Future perspectives

Epidemiology continues to face both opportunities and challenges. The potential access to big data provides opportunities (e.g., for artificial intelligence and machine learning), but with increased use of data that are not collected for non-research purposes it is likely that methodological obstacles such as confounding, missing data and measurement error are becoming more prevalent or more severe. It is sometimes claimed that data collected for research purposes do not reflect daily practice. It is important to recognise, however, that, conversely, evidence that originates from daily practice does not necessarily provide valid evidence for

daily practice. In the presence of difficult challenges, it is tempting to change one's inferential goals so that they become easier to achieve. However, this may leave the question that is of actual interest unanswered. If the interest is with a causal estimand, researchers should be explicit about this (Hernán, 2018).

Along with committing to a causal estimand, use of a causal roadmap may help avoid conflation of different parts of causal inference (Petersen and Van der Laan, 2014; Ahern, 2018). We believe that a distinction between identification and estimation is particularly useful as it means that the purely statistical issues of the latter can be put aside when concentrating on the former. At each step of the roadmap, there are areas for future methodological research.

For example, regarding missing data, emphasis is often placed on the classification of missingness as either being 'completely at random' (MCAR), 'at random' (MAR), or 'not at random' (NMAR), or on the recoverability of the entire joint distribution of a collection of variables. However, specific causal estimands might be identifiable even if the entire joint distribution cannot be recovered. For example, in case-control studies, the topic of **chapter 11**, certain causal effects may actually be identifiable from the observed data distribution while absolute risks are typically not.

When estimands are not identifiable, it may be possible to obtain partial identification bounds, which may preclude the estimand from taking, say, the null value of no causal effect. Partial identification is an interesting area for future research in part because it may inform sensitivity analyses.

Finally, rather than concentrating on methodological obstacles in isolation, we believe there may be value in considering multiple problems together (Van Smeden et al., 2021). After all, in applied research, epidemiologists often face multiple problems simultaneously and how they are best handled together is rarely obvious.

References

Ahern, J. (2018): "Start with the "C-word," follow the roadmap for causal inference," *American Journal of Public Health*, 108, 621.

Ali, M. S., R. H. Groenwold, S. V. Belitser, P. C. Souverein, E. Martín, N. M. Gatto, C. Huerta, H. Gardarsdottir, K. C. Roes, A. W. Hoes, et al. (2016): "Methodological comparison of marginal structural model, time-varying cox regression, and propensity score methods: the example of antidepressant use and the risk of hip fracture," *Pharmacoepidemiology and drug safety*, 25, 114–121.

Blakely, T., J. Lynch, and R. Bentley (2016): "Commentary: DAGs and the restricted potential outcomes approach are tools, not theories of causation," *International journal of epidemiology*, 45, 1835–1837.

Broadbent, A., J. P. Vandenbroucke, and N. Pearce (2016): "Response: formalism or pluralism? A reply to commentaries on 'Causality and causal inference in epidemiology'," *International Journal of Epidemiology*, 45, 1841–1851.

Daniel, R. M., B. L. De Stavola, and S. Vansteelandt (2016): "Commentary: The formal approach to quantitative causal inference in epidemiology: misguided or misrepresented?" *International journal of epidemiology*, 45, 1817–1829.

Hernán, M. A. (2018): "The C-word: scientific euphemisms do not improve causal inference from observational data," *American journal of public health*, 108, 616–619.

Krieger, N. and G. Davey Smith (2016): "The tale wagged by the dag: broadening the scope of causal inference and explanation for epidemiology," *International journal of epidemiology*, 45, 1787–1808.

Lau, B., P. Duggal, S. Ehrhardt, H. Armenian, C. C. Branas, G. A. Colditz, M. P. Fox, S. E. Hawes, J. He, A. Hofman, et al. (2020): "Perspectives on the future of epidemiology: a framework for training," *American journal of epidemiology*, 189, 634–639.

Lipsitch, M., E. T. Tchetgen, and T. Cohen (2010): "Negative controls: a tool for detecting confounding and bias in observational studies," *Epidemiology (Cambridge, Mass.)*, 21, 383.

Mitra, R. and J. P. Reiter (2016): "A comparison of two methods of estimating propensity scores after multiple imputation," *Statistical methods in medical research*, 25, 188–204.

Morris, T. P., I. R. White, and M. J. Crowther (2019): "Using simulation studies to evaluate statistical methods," *Statistics in medicine*, 38, 2074–2102.

Pearl, J. (2010): "On the consistency rule in causal inference: Axiom, definition, assumption, or theorem?" *Epidemiology*, 21.

Pearl, J., E. Bareinboim, et al. (2014): "External validity: From do-calculus to transportability across populations," *Statistical Science*, 29, 579–595.

Petersen, M. L. and M. J. Van der Laan (2014): "Causal models and learning from data: integrating causal modeling and statistical estimation," *Epidemiology (Cambridge, Mass.)*, 25, 418.

Richardson, T. S. and J. M. Robins (2013): "Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality," *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128, 2013.

Robins, J. M. (1999): "[Choice as an alternative to control in observational studies]: comment," *Statistical Science*, 14, 281–293.

Robins, J. M., T. J. VanderWeele, and R. D. Gill (2015): "A proof of bell's inequality in quantum mechanics using causal interactions," *Scandinavian Journal of Statistics*, 42, 329–335.

Robins, J. M. and M. B. Weissman (2016): "Commentary: counterfactual causation and streetlamps: what is to be done?" *International journal of epidemiology*, 45, 1830–1835.

Schwartz, S., S. J. Prins, U. B. Campbell, and N. M. Gatto (2016): "Is the "well-defined intervention assumption" politically conservative?" *Social science & medicine (1982)*, 166, 254.

Smeden, M., van, B. B. L. Penning de Vries, L. Nab, and R. H. H. Groenwold (2021): "Approaches to addressing missing values, measurement error, and confounding in epidemiologic studies" *Journal of Clinical Epidemiology*, 131, 89–100.

Souverein, P. C., V. Abbing-Karahagopian, E. Martin, C. Huerta, F. de Abajo, H. G. Leufkens, G. Candore, Y. Alvarez, J. Slattery, M. Miret, et al. (2016): "Understanding inconsistency in the results from observational pharmacoepidemiological studies: the case of antidepressant use and risk of hip/femur fractures," *pharmacoepidemiology and drug safety*, 25, 88–102.

Vandenbroucke, J. P., A. Broadbent, and N. Pearce (2016): "Causality and causal inference in epidemiology: the need for a pluralistic approach," *International journal of epidemiology*, 45, 1776–1786.

VanderWeele, T. J. (2016): "Commentary: on causes, causal inference, and potential outcomes," *International journal of epidemiology*, 45, 1809–1816.

VanderWeele, T. J., M. A. Hernán, E. J. Tchetgen Tchetgen, and J. M. Robins (2016): "Re: Causality and causal inference in epidemiology: the need for a pluralistic approach," *International journal of epidemiology*, 45, 2199–2200.

VanderWeele, T. J., A. R. Luedtke, M. J. van der Laan, and R. C. Kessler (2019): "Selecting optimal subgroups for treatment using many covariates," *Epidemiology (Cambridge, Mass.)*, 30, 334.