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

Penning de Vries, B.B.L.

1

General introduction and outline of thesis

What it means for something to cause something else has long been, and still is, a topic of debate among philosophers. In epidemiology and other sciences, a now-dominant approach to causality—the *counterfactual* or *potential outcomes framework*—considers causal claims that address what-if questions, claims about the consequences of hypothetical—possibly contrary-to-fact—actions (Neyman et al., 1935; Rubin, 1974; Holland, 1986, 1988; Pearl, 2009; Hernán and Robins, 2020). Statements of this form arise naturally in medicine, for example, where we are often faced with the problem of having to choose between treatment options for a patient and where—to guide decision making—we might ask what would happen with the patient's well-being if we choose one treatment option versus another.

1.1 Causal inference

Within the counterfactual outcomes framework, a *causal effect* is defined in terms of a contrast between at least two hypothetical actions. That at most one of mutually exclusive actions will actually become factual means that causal effects are not directly apparent. To proceed in our endeavour to quantify a causal effect, we start by looking for a way to equate our target quantity, the *estimand*, with a known function of the distribution of factual, real-world variables (Petersen and Van der Laan, 2014; Ahern, 2018). If we succeed in finding such a way, the estimand is said to be *identifiable* (from this distribution of factuals) and the expression that describes how it relates to the distribution is an *identifiability expression*. If the distribution of factuals is compatible with at least two values

of the estimand, the estimand is said to be non-identifiable and no identifiability expression exists.

The identifiability expression forms the basis for estimation. Where identification connects the estimand with the theoretical (sometimes called 'population') distribution of factual variables, estimation concerns how these theoretical distributions in turn relate to empirical distributions or finite random samples. An *estimator* is a function of a random sample that is intended to offer a close approximation of the estimand. Because the estimator is a function of a random sample, its output (sometimes also labeled the estimator) is itself a random variable. Given a fixed realisation of the sample, the output of an estimator is fixed and known as an *estimate*.

1.2 Obstacles in causal inference

There are a number of obstacles in providing 'good' estimates of causal effects. Some of these are purely statistical and relate only to estimation but others complicate identification as well as estimation. Three of these—confounding, missing data, and measurement error—are the main themes of this thesis and are described briefly below by way of an example.

Suppose that the interest lies with a contrast between someone's risk of sustaining a health-related event had this person been subjected, possibly contrary to fact, to exposure $A = 1$ (e.g., initiating a certain drug treatment, immediately after diagnosis with some condition) and the risk of the possibly contrary-to-fact situation where A was set to 0 (e.g., not initiating the drug treatment immediately after diagnosis). To clarify which hypothetical or counterfactual situation we are referring to, we write $Y(0)$ for the indicator of the event of interest that would be realised had the person been exposed to $A = 0$ and we likewise denote by $Y(1)$ the event indicator that would be realised had the person been exposed to $A = 1$. In the literature, and also in this thesis, rather than enclosing it within parentheses, a reference to a counterfactual situation is sometimes written in superscript (e.g., Y^0) or subscript (e.g., Y_0) to indicate the corresponding counterfactual version of a variable. In this example, the event indicators each take one of two levels, 1 in case the event of interest takes place and 0 otherwise. The interest can now be succinctly written as

$$\Pr(Y(1) = 1) \text{ versus } \Pr(Y(0) = 1). \tag{1.1}$$

1.2.1 Confounding

The simplest attempt at identifying the components of the contrast (1.1) is to replace the counterfactual event risk $\Pr(Y(a) = 1)$ with $\Pr(Y = 1|A = a)$ for $a = 0, 1$, yielding

$$\Pr(Y = 1|A = 1) \text{ versus } \Pr(Y = 1|A = 0). \tag{1.2}$$

The event indicator Y is the factual (observed), real-world outcome variable. But under what conditions are (1.1) and (1.2) actually equivalent, or when is $\Pr(Y(a) = 1)$ identified by $\Pr(Y = 1|A = a)$? Three key identifiability conditions are consistency, exchangeability, and positivity.

Consistency connects the counterfactual variables of interest with the factual, real-world variables (Cole and Frangakis, 2009; VanderWeele, 2009; Pearl, 2010). It means that the counterfactual version of an outcome variable (e.g., characterising the well-being of a patient) under a given hypothetical action coincides with its factual version if this hypothetical action agrees with (our impression of) the real world. For example, if a patient is known to have received treatment with a particular drug, consistency implies that the patient's well-being is the same as the patient's well-being that would have been realised had the patient been assigned this treatment. This seemingly trivial condition has two noteworthy subtleties. First, a prerequisite is that the hypothetical actions of interest are sufficiently well-defined. There may be many variations on "assigning drug treatment" (e.g., in the dosing or timing of the treatment) and their impact on the patient's well-being need not be the same. The actions are sufficiently well-defined if there is no ambiguity about the variation or all possible variations equally affect the outcome variables of interest (i.e., there is treatment variation irrelevance). Second, it may be that a patient's treatment is misclassified, resulting in a wrong impression about the real world. In turn, the patient's well-being need not coincide with its counterfactual counterpart that would be realised had the received the registered treatment. We will reconsider misclassification below. For now, let us assume that there is no such misclassification and that $Y(a) = Y$ if $A = a$, for $a = 0, 1$, so that

$$\begin{aligned}
\Pr(Y(a) = 1) &= \Pr(Y(a) = 1|A = a)\Pr(A = a) \\
&\quad + \Pr(Y(a) = 1|A = 1 - a)\Pr(A = 1 - a) \\
&\qquad\qquad \text{(by the law of total probability)} \\
&= \Pr(Y = 1|A = a)\Pr(A = a) \\
&\quad + \Pr(Y(a) = 1|A = 1 - a)\Pr(A = 1 - a). \quad \text{(by consistency)}
\end{aligned}$$

The application of the law of total probability above and (1.2) require that the conditional probabilities given $A = 1$ or $A = 0$ are defined, i.e., that $A = 1$ and $A = 0$ have positive probability (*positivity*).

Note that the right-hand side of the above equality has only one term that contains a counterfactual outcome. Under exchangeability, this term can be replaced and the above expression can be turned into an identifiability expression. *Exchangeability*, or 'no confounding', here means that $Y(a) \perp\!\!\!\perp A$—shorthand for $Y(a)$ is independent of A—for $a = 0, 1$, so that

$$\begin{aligned}\Pr(Y(a) = 1|A = 1 - a) &= \Pr(Y(a) = 1|A = a) && \text{(by exchangeability)}\\ &= \Pr(Y = 1|A = a). && \text{(by consistency)}\end{aligned}$$

Hence, under consistency, exchangeability and positivity,

$$\begin{aligned}\Pr(Y(a) = 1) &= \Pr(Y = 1|A = a)\Pr(A = a)\\ &\quad + \Pr(Y = 1|A = a)\Pr(A = 1 - a)\\ &= \Pr(Y = 1|A = a)[\Pr(A = a) + \Pr(A = 1 - a)]\\ &= \Pr(Y = 1|A = a)\end{aligned}$$

and, so, (1.1) and (1.2) are equivalent. When the exchangeability condition is violated, however, $\Pr(Y(a) = 1|A = a) \neq \Pr(Y(a) = 1|A = 1 - a)$ for $a = 0$ or $a = 1$, and therefore $\Pr(Y = 1|A = 0)$ and $\Pr(Y = 1|A = 1)$ might not equal $\Pr(Y(0) = 1)$ and $\Pr(Y(1) = 1)$, respectively.

Departure from identification of an estimand, brought about for example by a violation of the exchangeability condition, may have a knock-on effect on the properties of an estimator. Given a sample of n exposure-outcome pairs (A_i, Y_i), natural estimators of the components of the contrast (1.2) are

$$\sum_{i=1}^{n} \frac{A_i}{\sum_{j=1}^{n} A_j} Y_i \text{ and } \sum_{i=1}^{n} \frac{1 - A_i}{\sum_{j=1}^{n}(1 - A_j)} Y_i. \tag{1.3}$$

Suppose that the number of exposures is fixed at m (i.e., $\sum_{i=1}^{n} A_i = m$). If for $a = 0, 1$, $Y_i|A_i = a$ has the same distribution as $Y|A = a$, then

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} \frac{A_i}{\sum_{j=1}^{n} A_j} Y_i\right] &= \frac{1}{m}\sum_{i=1}^{n} \mathbb{E}[A_i Y_i]\\ &= \frac{1}{m}\sum_{i=1}^{n} \mathbb{E}[Y_i|A_i = 1]\Pr(A_i = 1)\\ &\quad \text{(by the law of total (or iterated) expectation)}\end{aligned}$$

$$
\begin{aligned}
&= \Pr(Y = 1|A = 1)\frac{1}{m}\sum_{i=1}^{n}\Pr(A_i = 1) \\
&= \Pr(Y = 1|A = 1)\frac{1}{m}\mathbb{E}\left[\sum_{i=1}^{n} A_i\right] \\
&= \Pr(Y = 1|A = 1).
\end{aligned}
$$

We therefore say that the estimators of (1.3) are unbiased for the components of (1.2) if $Y_i|A_i \sim Y|A$ and in turn for the components of (1.1) if additionally the identifiability assumptions (consistency, exchangeability and positivity) are met. More generally, given fixed θ, we say that an estimator $\hat{\theta}$ is unbiased for a quantity θ if $\mathbb{E}[\hat{\theta}] = \theta$, and we refer to $\mathbb{E}[\hat{\theta} - \theta]$ as the bias of the estimator $\hat{\theta}$ (relative to θ). This notion of bias relates to estimation. However, in this thesis we also occasionally use the term bias to refer to the distance between the quantity that is identified by the identification strategy and the value of the estimand. Of note, there is often a connection between the bias of an estimator and the bias of an identification strategy: in many cases, the former converges in some sense to the latter as the sample size on which the estimator is based increases. Confounding is often labeled a 'source of bias'—an apt description regardless of the notion of bias.

Bias, or the lack thereof, is only one of many properties of an estimator that describe how 'good' an estimator approximates a target quantity. The other properties include, but are not limited to, variance and mean squared error, which are addressed in other chapters of this thesis.

1.2.2 Missing data

Identification is a relative notion—it is relative to a set of factual variables whose distribution we seek to connect to the estimand. Including in this set of factuals variables that are not observed by the end of data collection has little practical value and, so, we eventually restrict our attention to the observed part. Before we do, however, it is often useful—although not always necessary—to consider, as an intermediate step, identification from a set of variables that may not be fully observed. If the distribution of these variables can be inferred (or 'recovered') from the distribution of the observed part, then identifiability from the former implies identifiability from the latter.

This insight motivates the classification of missingness as either *missingness* that is *completely at random* (MCAR), *at random* (MAR), or *not at random* (MNAR) (Rubin, 1976). Consider a sequence $(X_1, ..., X_p)$ of p variables and

an equally long sequence $(R_1, ..., R_p)$ of response indicators, with $R_i = 0$ if X_i is observed and $R_i = 1$ otherwise, $i = 1, ..., p$. Data are said to be MAR relative to a realisation $(r_1, ..., r_p)$ of $(R_1, ..., R_p)$ and realisations $(x_i : r_i = 1)$ of $(X_i : r_i = 1)$, if for all levels $(x_i : r_i = 0)$ of $(X_i : r_i = 0)$,

$$\begin{aligned}&\Pr(R_1 = r_1, ..., R_p = r_p | X_1 = x_1, ..., X_p = x_p)\\ &\quad = \Pr(R_1 = r_1, ..., R_p = r_p | X_i = x_i : r_i = 1);\end{aligned}$$

and MCAR relative to this realisation (a stronger condition) if

$$\Pr(R_1 = r_1, ..., R_p = r_p | X_1 = x_1, ..., X_p = x_p) = \Pr(R_1 = r_1, ..., R_p = r_p).$$

Missingness that is not 'at random' (or 'completely at random') is 'not at random'. An interesting special case where MAR is always satisfied is the case where the first j variables, $1 \leq j \leq p$, are always observed (i.e., $R_1 = ... = R_j = 1$) and the missingness of the other $p - j$ variables satisfies

$$(R_i : j < i \leq p) \perp\!\!\!\perp (X_i : j < i \leq p) | (X_i : 1 \leq i \leq j).$$

For this special MAR condition, it is easy to determine whether the joint distribution of $(X_1, ..., X_p, R_1, ..., R_p)$ is recoverable from the distribution of its observed part: the distribution of the partially observed (i.e., last $p-j$) variables given the always observed (i.e., first j) variables is obtained by conditioning on $R_1 = ... = R_p = 1$.

It is interesting to consider counterfactual outcomes that are strictly 'non-factual' as missing variables. If $Y(0) = Y$ whenever $A = 0$ and there are no missing factuals, the missingness of $Y(0)$ in (1.1) is fully determined by A: $Y(0)$ is observed if and only if $A = 0$. The missingness is therefore 'at random'. The counterfactual outcome probability $\Pr(Y(0) = 1 | A = 1)$ is however not identified by $\Pr(Y(0) = 1 | A = 1, R_0 = 1)$, with R_0 denoting the response indicator for $Y(0)$, because $R_0 = 1$ if and only if $A = 0$ and, since $\Pr(A = 1, A = 0) = 0$, $\Pr(Y(0) = 1 | A = 1, A = 0)$ is not defined.

1.2.3 Measurement error

Measurement error arises when our impression of a variable's value is different from its actual value. The variable can be considered unobserved and its potentially wrong impression forms another, observed variable.

A special form of measurement error is misclassification, which means that our measurement or impression of a categorical variable is inaccurate. As for the above example, exposure misclassification means that our measurement or

'impression' A^* of A does not always coincide with A itself. While $Y(a)$ might equal Y if $A = a$ for $a = 0, 1$, it is possible that $Y(a) \neq Y = Y(A)$ if $A^* = a$. This is sometimes described as a possible violation of the consistency assumption (Gravel and Platt, 2018); the counterfactual outcome $Y(a)$ need not coincide with the observed Y even if $A^* = a$. Upon replacing A with A^*, (1.2) becomes

$$\Pr(Y = 1|A^* = 1) \text{ versus } \Pr(Y = 1|A^* = 0). \tag{1.4}$$

A simple yet naïve approach is to take (1.4) as the basis for inference about (1.1). However, that (1.1) and (1.4) are equivalent is not evident and may not be true. More generally, like confounding and missing data, measurement error is an obstacle in causal inference.

1.3 Objective and outline of thesis

There are many approaches to handling the obstacles of confounding, missing data and measurement error. To address confounding, these approaches include traditional regression analyses and the more modern propensity score methods such as propensity score matching and inverse probability weighting (Rosenbaum and Rubin, 1983; Robins et al., 2000). These methods rely on the availability of other variables, covariates, such that conditional on these covariates, there is exchangeability. Other methods, like instrumental variable analysis and negative control methods, rest on different assumptions (Greenland, 2000; Lipsitch et al., 2010). For missing data, simply discarding incomplete records has long been the default approach. More principled approaches include expectation-maximisation, multiple imputation, and inverse probability weighting (Dempster et al., 1977; Rubin, 1987; Robins et al., 2000). Lastly, the impact of measurement error may be mitigated by regression calibration, simulation extrapolation (SIMEX), latent variable modelling, or inverse probability weighting (Buonaccorsi, 2010; Gravel and Platt, 2018).

The development and study of the properties of methods to overcome the abovementioned methodological obstacles in epidemiology is an active area of research with many open questions. The aim of this thesis is to contribute to this research and to provide more insight into the properties of methods for confounding, missing data and measurement error.

The outline for the remainder of this thesis is as follows. We start by considering in **chapter 2** the task of reviewing the existing literature to gauge the current state of knowledge about existing methodologies, identify gaps or to provide a starting point for guidance development. The chapter gives an

appraisal of methods for gathering information about the use of methods for the study of causal effects. In all subsequent chapters, we zoom in on methods of the latter kind. In **chapters 3 and 4**, we address the concern that the combination of multiple imputation for missing data and propensity score methods for confounding has worse performance than might be expected from how they perform in isolation. In doing so, we compare two strategies from epidemiological practice for implementing propensity score methods to multiply imputed data sets and we give guidance on which is to be preferred. The study of propensity score methods and missing data is continued in **chapter 5**, which focuses on a class of machine learning methods, classification and regression trees (CART), for estimating propensity scores in the presence of missing covariate data. **Chapter 6** then turns to propensity score matching and missing outcome data. The chapter illustrates that when baseline exchangeability is achieved through propensity score matching, bias might result from restricting downstream analysis to the subset of individuals who have not dropped out of the study by the administrative study end. In **chapter 7**, we consider missing data mechanisms that are governed by study design. In studies on the effects of time-varying exposures, adequate information on time-varying participant characteristics might help mitigate time-dependent confounding. However, the frequency with which these characteristics are measured may be inadequate and participant characteristics are sometimes (wrongly) assumed to remain constant in periods of no measurement (i.e., there is measurement error). The chapter illustrates the impact of design choices regarding data collection. Measurement error is also a dominant theme in **chapter 8**, in which a weighting method for simultaneous adjustment for confounding and joint exposure-outcome misclassification is developed. The method relies on standard identifiability assumptions, such as exchangeability within levels of a collection of partially observed variables, consistency, positivity, and MAR. In many studies on causal effects, however, there are often concerns that standard identifiability assumptions are violated. Negative controls are a tool with the potential to detect or correct for confounding that is explained by fully unobserved variables. This is the topic of **chapter 9**. The counterfactual outcomes framework and attempts to identify estimands have become increasingly popular in much of the causal inference literature, including the literature on negative controls. Case-control studies have not yet enjoyed this trend. In **chapter 10**, we reconsider this family of designs and recast classical concepts, assumptions and principles from a modern perspective. It is shown how and when a variety of causal estimands can be identified with these study designs. Causal inference and prediction are two areas of epidemiology that are increasingly seen as overlapping. In precision

medicine, it is not uncommon for treatment assignment decisions to be based on 'prognostic scores', predictions of the outcome of interest that would be realised if the treatment were withheld. **Chapter 11** deals with this topic and emphasises that in order to obtain optimal results, the counterfactual outcomes under both treatment levels, 'treatment' and 'no treatment', should be considered. The methodological obstacles that we encounter in causal inference, including confounding, missing data and measurement error, are therefore relevant in that context too. To conclude, in **chapter 12**, we present a summary of the previous chapters, along with a general discussion of this thesis in the light of the existing literature, with suggestions for future research.

1.4 References

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

Buonaccorsi, J. P. (2010): *Measurement error: models, methods, and applications*, Boca Raton: Chapman & Hall/CRC.

Cole, S. R. and C. E. Frangakis (2009): "The consistency statement in causal inference: a definition or an assumption?" *Epidemiology*, 20, 3–5.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 1–22.

Gravel, C. A. and R. W. Platt (2018): "Weighted estimation for confounded binary outcomes subject to misclassification," *Statistics in medicine*, 37, 425–436.

Greenland, S. (2000): "An introduction to instrumental variables for epidemiologists," *International journal of epidemiology*, 29, 722–729.

Hernán, M. and J. Robins (2020): *Causal Inference: What If*, Boca Raton: Chapman & Hall/CRC.

Holland, P. (1986): "Statistics in causal inference," *Journal of the American Statistical Association*, 81, 945–960.

Holland, P. (1988): "Causal inference, path analysis, and recursive structural equations models," *Sociological Methodology*, 18, 449–484.

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.

Neyman, J., K. Iwaszkiewicz, and St. Kolodziejczyk (1935): "Statistical problems in agricultural experimentation," *Supplement to the Journal of the Royal Statistical Society*, 2, 107–180.

Pearl, J. (2009): *Causality: Models, Reasoning and Inference*, New York: Cambridge University Press.

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

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, 418–426.

Robins, J. M., M. A. Hernan, and B. Brumback (2000): "Marginal structural models and causal inference in epidemiology," *Epidemiology*, 11.

Rosenbaum, P. R. and D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rubin, D. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688–701.

Rubin, D. B. (1976): "Inference and missing data," *Biometrika*, 63, 581–592.

Rubin, D. B. (1987): *Multiple imputation for survey nonresponse*, New York: Wiley.

VanderWeele, T. J. (2009): "Concerning the consistency assumption in causal inference," *Epidemiology*, 20, 880–883.