Causal inference: Critical developments, past and future

Erica E. M. MOODIE1* and David A. STEPHENS2

1Department of Epidemiology and Biostatistics, McGill University, 2001 McGill College Ave, Montréal, Quebec, Canada H3A 1G1
2Department of Mathematics and Statistics, McGill University, 805 Sherbrooke St W, Montréal, Quebec, Canada H3A 2K6

Key words and phrases: Counterfactuals; potential outcomes; propensity scores; semiparametric estimation.

MSC 2020: Primary 62A01; secondary 62H20.

Abstract: Causality is a subject of philosophical debate and a central scientific issue with a long history. In the statistical domain, the study of cause and effect based on the notion of “fairness” in comparisons dates back several hundreds of years, yet statistical concepts and developments that form the area of causal inference are only decades old. In this article, we review the core tenets and methods of causal inference and key developments in the history of the field. We highlight connections with traditional “associational” statistical methods, including estimating equations and semiparametric theory, and point to current topics of active research in this crucial area of our field.

Résumé: La causalité est un sujet de débat philosophique et une question scientifique fondamentale ayant une longue histoire. En statistique, l’étude des concepts de causes et effets fondée sur la notion d’«équité» dans les comparaisons remonte à plusieurs centaines d’années, alors que les concepts et les développements statistiques qui composent le domaine de l’influence causale ne datent que de quelques décennies. Cet article examine les principes et les méthodes de base de l’influence causale ainsi que l’historique de leurs développements majeurs. Nous y mettons en évidence leurs liens avec les méthodes statistiques «associationnelles» traditionnelles, notamment les équations d’estimation et la théorie semi-paramétrique, et soulignons les sujets de recherche actifs dans ce domaine critique de notre domaine.

1. INTRODUCTION

In this article, we review the rapidly developing literature on causal inference from a statistical perspective. Causal thinking has long guided study design and analysis, but only relatively recently have the core ideas come into the statistical mainstream. Alongside the development of paradigms for expressing priorities in a causal setting, statistical frameworks and methods have been the focus of much research.

A central objective in many statistical settings is to infer (assess, test for) the impact of an intervention on an outcome on the basis of study data. The central challenge is that, typically, the data available are not suitable for making such an assessment without further assumptions or adjustments because, although the intervention can be observed in the data, variation in the
intervention occurs in an uncontrolled manner and it is plausible that any observed effect may be
due to other factors. Fundamentally, it is this “confounding” of the intervention effect that must
be accounted for using statistical techniques.

In modern statistical research and related areas of machine learning, there is considerable
mystique and many misconceptions surrounding the nature of causal inference. As we will
discuss, it is false to suppose that any analysis simply using, say, a model for treatment allocation
(a “propensity score”) will return a causal estimate. At the same time, it is equally false to suggest
that ordinary conditional regression models cannot provide causal estimates: although there are
some settings in which this is true, it is not universally the case.

Some of the most prominent and important leaders in the field of causal inference have, unfor-
tunately, contributed to these views. For example, in an introductory paper on causal inference,
Pearl (2010) writes “causal and associational concepts do not mix” and follows this by stating

A useful demarcation line between associational and causal concepts crisp and easy to
apply, can be formulated as follows. An associational concept is any relationship that can be
defined in terms of a joint distribution of observed variables, and a causal concept is any
relationship that cannot be defined from the distribution alone.

Whereas this point is well taken from a philosophical perspective, in statistical terms it
creates a false dichotomy, one that regretfully often serves to create tension where none need
exist. Traditional, nonpredictive statistical methods typically do aim to answer causal questions.
The core tenet of causal inference is to do so through careful consideration of context and
through explicit statements of assumptions regarding the underlying data-generating mechanism
(“nature”) as well as the estimands (what contrasts, applied to which population).

2. CORE IDEAS

Causal inference typically relies on several identification assumptions that correspond to implicit
assumptions made in traditional regression analyses, as we shall detail below. The formulation of
causal inference problems, however, begins several steps earlier with overt focus on estimands,
or target parameters of interest, which are defined in terms of contrasts between treatment levels
within a specific population.

A causal approach to an analysis must first define the causal research question and then
typically proceeds as follows (Goetghebeur et al., 2020):

1. Formalize all definitions:
   1.1. define the treatment and its relevant levels;
   1.2. define the outcome, specifying, if necessary, the time following treatment at which it
        is measured; and
   1.3. define the population(s) of interest.

2. Specify the estimand (target causal effect).

3. Perform the estimation:
   3.1. state identifiability and estimation assumptions;
   3.2. estimate the causal effect; and
   3.3. evaluate assumptions and perform sensitivity analyses.

4. Perform inference.

Causal estimands (step 2) are often formulated in terms of potential outcomes, also known as
counterfactuals. A critical component to determining assumptions (step 3.1) requires the analyst
to determine via existing literature or substantive experts the relevant variables needed to posit
a realistic model for the data-generating structure. This may be formalized in a directed acyclic graph (DAG), as will be detailed below. Estimating a causal effect can often be accomplished using a variety of approaches, which include both traditional statistical methods and those considered to be “causal approaches” such as those based on the propensity score. Some variations on the above roadmap have been proposed (e.g., Shimoni et al., 2019), although a focus on formalizing and making explicit all assumptions is common.

The final step, of performing inference, is afflicted by the same challenges in the causal framework as in standard statistical settings. Calculating measures of variability or (un)certainty such as standard errors or confidence intervals is often computationally straightforward. As we shall see in Section 4.3, many causal methods of estimation require two-step estimators that rely on plug-in estimates or call for the marginalization of conditional models. Asymptotically, many such methods do not require a correction for the substitution estimator (Henmi & Eguchi, 2004); nonetheless, a standard nonparametric bootstrap is commonly employed to guard against optimism in variance calculations that ignore nuisance parameters in finite samples. There the simplicity ends, as decision making—drawing conclusions about the “significance” of a finding—is no less controversial in the causal realm than in standard (frequentist) statistics (Greenland & Rafi, 2020).

We denote an intervention or treatment by $Z$, pretreatment covariates by $X$, and observed outcomes by $Y$. We denote potential outcomes, defined in the following section, by $Y(z)$ for all $z$ in some treatment set $\mathcal{Z}$: loosely, we can consider the potential outcome $Y(z)$ to be the random variable that records the outcome that would occur (perhaps contrary to what is observed) if the treatment was set to $z$ through the intervention of the researcher. In, say, agricultural and medical settings, the concept of a controlled experiment where intervention occurs by design is long established and is the natural setting within which to initiate a consideration of causal concepts. We expand on this in the following section.

3. KEY HISTORICAL DEVELOPMENTS

3.1. Randomization and Confounding

Though perhaps only really recognized in hindsight (Rubin, 1990), Neyman is now credited with being the first to describe the notion of a potential outcome (Neyman, 1923) when he described the (unknown) “potential” yield of plots under varying conditions. This concept of potential outcomes has been used to formulate an entire framework (Rubin, 1974) on which much of modern causal inference is based. However, notions of causality in statistics, particularly medical statistics, arose earlier.

The first known clinical trial is often attributed to James Lind’s 1747 study of citrus to treat scurvy (Upadhyayula & Kasliwal, 2020). The first randomized clinical trial occurred much later (Medical Research Council, 1949) following principles laid out by Sir Austin Bradford Hill (1937) in the previous decade, which were built on the design of experiments of Sir Ronald Aylmer Fisher from the 1920s and beyond; see e.g., Fisher (1951) or, for an excellent summary, Fisher Box (1980). While these works are foundational and formalize statistical issues around the notion of confounding, the idea was not entirely new: recent work (Chalmers, 2011) has highlighted notions of “fairness” that date back to the 14th century (Petrarca, 1364):

I solemnly affirm and believe, if a hundred or a thousand men of the same age, same temperament and habits, together with the same surroundings, were attacked at the same time by the same disease, that if one half followed the prescriptions of the doctors of the variety of those practising at the present day, and that the other half took no medicine but relied on Nature’s instincts, I have no doubt as to which half would escape.

It is this concept of fairness in comparison groups, or the lack of confounding, that is central to causal comparisons. Randomization aims to find groups of individuals whose covariates (e.g.,
the age, temperament, and habits noted by Petrarca) follow the same distribution and allow only one factor—the treatment of interest—to vary systematically. Potential outcomes provide a notation and a means of expressing this at an individual level. That is, for a binary treatment $Z \in \{0, 1\}$, we may consider the outcomes of an individual—whose covariates are fixed and unchanging—under the two values of treatment. We denote these random variables by $Y(0)$ and $Y(1)$, respectively. Potential outcomes with individual index $i$ have variably been denoted by $Y_i(z)$, $Y_{i}^z$, or by distributional equivalents such as $P(Y = y|\text{set}(Z = z))$ or $P(Y = y|\text{do}(Z = z))$, where the latter two form the basis of what is known as Pearl’s “do-calculus” (Pearl, 2009) and can be used to consider outcome distributions or estimands based on, say, expectations of these distributions at the population level. Extensions of potential outcomes to multivalued or continuous treatments are straightforward.

### 3.2. Towards Drawing Causal Conclusions

Prior to the formalization of causal inference via the potential outcomes framework, however, rather less rigorous but important discussions of causality arose in the field of medical statistics. Perhaps the best known ideas are what are often referred to as the “Bradford Hill criteria” proposed in “The Environment and Disease: Association or Causation?” (Bradford Hill, 1965): strength of relationship, consistency, specificity, temporality (cause precedes effect), biological gradient, plausibility, coherence, experiment (trials), and analogy.

Bradford Hill specifically noted that these are not criteria, but rather a group of conditions that might be useful to assess (rather than establish) causality. In an earlier lecture, Bradford Hill (1962) asked

> The question, on the other hand, may well be asked, what does one accept as overwhelming?

and later stated

> We are continuously brought back to the fundamental question—what alternative explanation will fit a set of observations, what other differences between our contrasted groups could equally, or better, account for the observed incidences.

The answer lies, Bradford Hill suggests, at least in part in a thorough understanding of the context and the science of the substantive question to be answered and “is not to be found tidily tucked up in the formulae of tests of significance, useful as they may be.”

Bradford Hill’s 1965 article was reprinted in 2020 in *Observational Studies*. The 1962 lecture was reprinted in *Statistics in Medicine*. Each was followed by commentary and discussion (Bradford Hill, 2020; Bradford Hill, 2022). These reflections highlight the many ways in which Bradford Hill was a visionary statistician and are a call to reflect on the manners in which current data and research questions may require different or additional assumptions and considerations.

### 3.3. Designing Observational Studies to Permit Causal Conclusions

In a seminal paper, Rubin (1974) described a variety of strategies that could be used to estimate causal effects from observational (nonrandomized) studies. This article was a rebuttal to the then prevailing wisdom that causal effects could be learned only from randomized studies. As Rubin noted, there are many instances where randomization is infeasible (e.g., due to cost or a long latency period of the exposure) or unethical (e.g., the exposure of interest is harmful). Rubin (1974) began by defining an average causal (or treatment) effect (ATE) as

$$E[Y_i(1) - Y_i(0)],$$

where, as above, $Y_i(z)$ for $z \in \{0, 1\}$ denotes the potential outcome for individual $i$ if they were to receive treatment $z$. Noting that both these potential outcomes can never be observed, Rubin
remarked that $Y_i(z)$ may serve as a good stand-in for $Y_j(z)$ if individuals $i$ and $j$ are “perfectly matched.” If the entire sample consists of $n$ such matched pairs, then the difference in the sample average among those treated and those not can estimate the ATE. Even if the matches are imperfect, the result should be close. Drawing an analogy between randomized and observational studies, Rubin noted that chance imbalance can occur even under randomization and “we cannot be guided solely by the concept of unbiasedness over the randomization set.” He further stated “We need some model for the effect of prior variables in order to use their values in an intelligent manner.” A concurrent publication (Cochran & Rubin, 1974) points to regression and matching “as two such means of addressing bias in observational studies.

Over the next several years, Rubin (e.g., Rubin, 1977, Rubin, 1979) produced a series of papers that focused primarily on matching and regression to remove bias in estimation due to measured confounding variables in observational studies and laid the foundations for the ground-breaking development of the propensity score (Rosenbaum & Rubin, 1983). A balancing score is a function of (pretreatment) covariates such that, conditional upon it, the treatment and those covariates are independent. The propensity score is the coarsest of all balancing scores. More formally, a balancing score $b(x)$ is any function of covariates $x$ such that $Z \perp X|b(x)$. Trivially, $x$ is a balancing score. For a balancing score to be of any use, the covariates $X$ on which balance is induced should include all relevant confounders (or at least the subset of all confounders needed to ensure that no spurious association between $Z$ and $Y$ may be found). See Section 4.1 for further discussion.

Define the propensity score by $e(x) = \Pr(Z = 1|x)$. It follows (Rosenbaum & Rubin, 1983) that $Z \perp X|e(x)$ and, further, that $(Y(0), Y(1)) \perp Z|e(x)$, which implies strong ignorability given the propensity score. This result may be used to establish that the ATE for a continuous outcome $Y$ can be found by (i) constructing a matched sample using the (univariate) propensity score rather than the typically higher dimensional covariate vector, (ii) subclassifying or stratifying based on the propensity score, or (iii) applying a linear regression model in which the balancing score is included as a covariate, in what is known as “propensity score regression.” The latter approach encompasses—under certain assumptions—traditional regression analysis because the full vector $x$ is a balancing score; see Section 4.3.1. Rosenbaum & Rubin (1983) further demonstrated that an estimated propensity score retains the desired balancing properties in finite samples provided the estimated propensity score is bounded away from 0 and 1.

In parallel to the literature on adjustment for confounding, a growing body of work on missing data emerged, partly reflecting the fact that potential outcomes themselves are largely unobserved. Robins, Rotnitzky & Zhao (1995) proposed semiparametric estimators for missing data known as “inverse probability weighted estimators,” where the probability that an observation was recorded conditional on measured covariates is used as a weight. Under the setting in which data are missing at random (Rubin, 1976) and missingness probabilities are bounded away from 0 and 1, the derived estimators are asymptotically consistent and semiparametric efficient. From this point, the connection to confounding was drawn: as Holland (1986) previously noted, the “fundamental problem” in causal inference is the impossibility of observing both $Y(0)$ and $Y(1)$ for any individual. This provided a lens through which to view potential outcomes as missing data, and led to the development of inverse probability of treatment weighting as a means of addressing confounding to estimate parameters from marginal models (Robins, 1998).

Through the 1990s and beyond, both methods of analysis and, importantly, designs of observational studies continued to be developed with the objective of reducing or eliminating bias due to confounding. For example, Rubin (2007) proposed the construction of an analytic dataset in an observational context such that treatment groups have balanced (comparable) covariate distributions without any knowledge of the outcome variable. Rubin’s approach aimed to emulate a randomized trial using techniques such as matching. These ideas continue to be refined to meet the challenges of ever-growing data sources, many of which produce data that are
not only observational but also not collected for research purposes (e.g., Hernán & Robins, 2016; Danaei et al., 2018; Murray, Marshall & Buchanan, 2021).

3.4. When Tradition Fails (Almost)

As we will demonstrate in Section 4.3, there are many estimands for which traditional statistical approaches such as regression adjustment can—under assumptions of correct model specification—yield consistent estimators. This includes the average treatment effect, the causal risk ratio, and so on. However, in the last few decades, the literature on causal inference has produced some estimands that are difficult to define without the aid of potential outcomes, even though it is possible to estimate these quantities using some combination or repeated application of traditional methods such as regression, prediction, and averaging.

Consider, for example, the average treatment effect on the treated (ATT): \( E[Y_i(1) - Y_i(0)] \mid Z = 1 \) (Imbens & Angrist, 1994). This estimand captures the effect of treatment in the population who were in fact treated, a population whose covariates typically differ from those of the population as a whole and for whom the treatment effect may be most relevant. For instance, if the exposure of interest is a nicotine substitution treatment, the impact of this treatment on the number of cigarettes smoked is not relevant to the whole population or perhaps even to the population of smokers. Instead, the causal effect of the nicotine substitution among those who are willing to accept treatment—and therefore are showing an interest or commitment in giving up smoking—is arguably the most relevant target population. While it is possible to estimate this quantity using regression and prediction, it is more commonly and more easily estimated using propensity score methods. Arguably, this estimand is less easily formulated without “causal thinking”: potential outcomes are not necessary to define the ATT, but they prove useful because the expectation \( E[Y_i(0)] \mid Z = 1 \) is not commonly considered in probabilistic modelling.

Another instance in which potential outcomes proved useful in the formulation of an important and meaningful outcome came with the introduction of marginal structural models (Robins, 1999a,b; Robins, Hernán & Brumback, 2000). Marginal structural models are typically defined in a setting where there is a longitudinal sequence of treatments, say, \( Z_1 \) and \( Z_2 \), and interest lies in the impact of that sequence of treatments on some final outcome \( Y \) through \( E[Y_i(z_1, z_2)] \) or a contrast between this expectation for a given pair of treatment combinations, e.g., \( E[Y_i(1,1) - Y_i(0,0)] \). This model is marginal rather than conditional in that it marginalizes over pretreatment covariates, including any intermediate variables that might be affected by \( Z_1 \) but precede \( Z_2 \). The term “structural” is used to indicate that the model is assumed to be causal, with estimands encompassing effects that can be derived from the true data-generating relationships. It is now well known that where an intermediate variable is both a mediator of the relationship between \( Z_1 \) and \( Y \) and a confounder of the relationship between \( Z_2 \) and \( Y \), there is no regression model parameter that corresponds to \( E[Y_i(z_1, z_2)] \) (e.g., Hernán, Brumback & Robins, 2000; Moodie & Stephens, 2010); a model that omits the intermediate variable will not appropriately adjust for confounding of the effect of \( Z_2 \) and yet including that variable in the model will bias the estimator of the effect of \( Z_1 \) by blocking any part of its impact that is mediated through that variable.

As an aside, we note that while marginal structural models are most often used in the context of time-varying exposures, these models can also be used in point-treatment settings. In fact, the ATE \( E[Y(1) - Y(0)] \) is a marginal structural model that marginalizes over all pretreatment covariates. Marginal models may also be only “partially” marginal; for example, it is possible to estimate a treatment effect that conditions on some particular covariate (say, older age or a specific comorbidity) and marginalizes over all other covariates.

In this section, we have provided a partial and brief survey of developments in statistical causal inference. We now turn to a formalization of the statistical assumptions as well as some typical methods employed to estimate quantities such as the ATE.
4. ASSUMPTIONS AND ESTIMATION

In this section, we formalize the tools and assumptions of causal inference, and provide a more technical description of the estimation approaches outlined in Sections 3.3 and 3.4.

4.1. Causal Graphs

DAGs, often called causal graphs, are a tool used to visually display causal assumptions made about the data-generating relationships in a given substantive problem and to assist in determining those variables that can distort or bias the causal effect(s) of interest (Pearl, 1995). A graph is said to be directed if all intervariable relationships are depicted by arrows, where an arrow from one variable into another indicates that the first variable causes changes in the second. The graph is said to be acyclic if it has no closed loops. Finally, for a DAG to be causal, all common causes of the treatment \( Z \) and the outcome \( Y \) (measured or otherwise) must be included in the graph (VanderWeele, Hernán & Robins, 2008).

Consider, for example, the left panel of Figure 1. In this very simple DAG, we see that \( X \) is a common cause of (and therefore a confounder of the relationship between) \( Z \) and \( Y \), while \( M \) is a mediator of the relationship between \( Z \) and \( Y \). DAGs can be elaborated to include more covariates, e.g., by considering confounders separately or by including variables that may affect \( Y \) but not \( Z \), or can be used to depict treatment sequences where treatments and intervening variables are observed at multiple time points, as in the right panel of Figure 1. The aim of causal inference is to quantify the “flow” of the effect of an intervention on treatment \( Z \) through all the directed paths that connect \( Z \) to \( Y \) while accounting for all the other paths that connect the two variables. For example, in the left panel of Figure 1, there are two paths by which \( Z \) may act on \( Y \) (one direct \( Z \rightarrow Y \) and the other passing through \( M \)). In contrast, there is an “open,” undirected path \( Z \leftarrow X \rightarrow Y \) through which a covariation in \( Z \) and \( Y \) may be observed unless the path is blocked. The path \( Z \leftarrow X \rightarrow Y \) is termed a confounding or backdoor path. Conditioning on \( X \) removes the confounding by blocking this path. For an excellent introduction to the use of causal DAGs to identify variables needed to block spurious paths between \( Z \) and \( Y \), we refer the interested reader to Greenland, Pearl & Robins (1999) or Didelez (2018). In the right panel of Figure 1, the situation is more complex, and conditioning strategies to remove confounding are less obvious (Moodie & Stephens, 2010).

Figure 2 illustrates the role of the propensity score in adjustment. The propensity score \( e(X) \) is a deterministic function of the confounders \( X \). Knowledge of \( e(X) \) determines the distribution of \( Z \) and renders \( Z \) conditionally independent of \( X \). Recall that \( e(X) \) is univariate irrespective of the dimension of \( X \).

Causal graphs are useful tools for displaying and clarifying proposed relationships between variables and for formulating causal questions. However, whereas DAGs encode specific probabilistic relationships, they do not uniquely define joint probability distributions. This

**Figure 1**: Examples of DAGs. The left panel shows a simple setting with a single treatment \( Z \) whose effect on the outcome \( Y \) is confounded by \( X \) and mediated through another variable \( M \). The right panel is a more complex longitudinal example with two treatments, \( Z_1 \) and \( Z_2 \), that occur in sequence and are temporally preceded by covariates \( X_1 \) and \( X_2 \), and followed by a final outcome \( Y \). A possibly unobserved variable \( U \) causally affects \( X_1, X_2, \) and \( Y \).
supports the comment of Pearl (2010) mentioned in the introduction. Consequently, learning
causal structure from observed data is a very challenging problem, which reinforces the
importance of the strategies outlined in Section 2 and Bradford Hill’s pragmatic formulation.

4.2. Identifiability Assumptions

Several key assumptions are made in much of the causal literature. While at first these may
appear to be restrictive, we shall see that, in fact, these assumptions are often made implicitly in
standard statistical inference.

Consistency, also sometimes referred to as the well-defined treatment assumption or treatment
variation irrelevance, states that any variation in the exposure of interest has no relevance to its
impact on the outcome. This could be true in a setting where the treatment is a dietary supplement
given either as a capsule or in drops, but is unlikely to hold when the “treatment” is “5-kg weight
loss” to be accomplished by exercise, caloric intake restriction, or some combination of these.
While weight loss caused by both exercise and caloric intake restriction may lead to reduced joint
pain, the impact of the method of weight loss on outcomes such as mood or overall self-reported
well-being may differ. This assumption is primarily made to ensure a “clean,” well-defined,
and reproducible result rather than for statistical purposes. This assumption is rarely if ever
stated in traditional statistical inference, yet few would disagree with its utility in terms of
making causal statements. Under consistency, potential outcomes are linked to observed data via
\[ Y = Y(0)(1 - Z) + Y(1)Z. \]

One independence assumption commonly made in the causal inference literature is known as
the stable unit treatment value assumption (SUTVA). This requires consistency and, further, states
that an individual’s outcome is affected only by the treatment they receive but not that of others
(Cox, 1958; Rubin, 1980). This is also referred to as a lack of interference or spill-over. SUTVA
is most easily explained in terms of examples where it does not hold. Consider, for example,
the case of a treatment by vaccination: the vaccination status of the members of one’s household
(workplace, community, etc.) has a significant bearing on an individual’s probability of encounter-
ing and hence catching an illness. Thus, the outcome for an individual is affected by the treatment
status of other individuals. SUTVA is thus unlikely to hold when the treatment is a vaccine.

The assumption of no unmeasured confounding is perhaps the strongest and also the least (sta-
tistically) controversial. Stated in terms of potential outcomes, it requires that \( \{Y(0), Y(1)\} \perp Z|X \).
That is, given the information contained in the covariates \( X \), the treatment received carries no
additional information on the individual’s potential outcomes. This assumption is also referred
to as conditional exchangeability or strong ignorability (Rosenbaum & Rubin, 1983). In more
traditional statistical terms, this is equivalent to saying that treatment is randomized within strata
defined by \( X \). Other forms of this assumption exist; for example, weak ignorability requires that
\( Y(z) \perp Z|X \) for each potential outcome \( Y(z) \) but not, as in strong ignorability, jointly for both
potential outcomes (Greenland & Robins, 2009).

Finally, the assumption of (conditional) positivity or overlap states that there exists no stratum
of \( X \) such that treatment assignment is uniquely determined (or where some levels of treatment
are impossible): \( P(Z = z | X = x) > 0 \) for all \( z \) and \( x \). This assumption is often stated in the context of analyses that rely on the propensity score and is a necessary assumption for nonparametric estimation of the ATE. However, positivity can be viewed through the lens of extrapolation in a more classical estimation setting. Suppose that, for a given value \( x^* \) of \( X \), \( P(Z = 0 | X = x^*) = 1 \). It is clear that there will be no triples of observed data \( (Y, Z, X) \) such that \( Z = 1 \) and \( X = x^* \); thus, any estimates of, say, a conditional ATE such as \( E[Y|Z = 1, X = x^*] - E[Y|Z = 0, X = x^*] \) would require extrapolation. Of course, there may be instances where we are willing to extrapolate, e.g., if we are willing to assume correct specification of the conditional mean model for \( Y \). If such a model cannot be assumed to be correct (either locally over the region where all pairs \((z, x)\) are observed or more globally), then positivity is required.

### 4.3. Commonly Employed Approaches to Estimating Causal Effects

Following a careful review of existing literature and consultation with subject-matter experts, an analyst should identify a sufficient set of confounders \( X \), possibly with the assistance of a DAG. Assuming that these confounders are all measured without error, the causal effect of a treatment can be estimated by comparing observed outcomes between different levels of the treatment between groups of individuals with identical or very similar values of \( X \). This comparison can be performed in a variety of ways including via regression, stratification, matching, or weighting.

#### 4.3.1. Traditional outcome regression modelling

Regression modelling can be used to undertake a causal analysis. A model must be specified for the outcome as a function of the treatment \( Z \) and the confounders \( X \). For instance, one could suppose a linear model such as

\[
E[Y|Z, X, \beta] = \beta_0 + \beta_Z Z + \beta_X^T X,
\]

when the outcome is continuous. The ordinary least squares estimator of \( \beta_Z \) is a consistent estimator of the ATE provided that the above model is correctly specified (i.e., the expected value of \( Y \) depends linearly on \( Z \) and \( X \), and there are no interactions between these variables). Alternatively, a more complex model could be fit to allow for, say, polynomial terms and treatment–covariate interactions. In the latter case, obtaining an estimator of the ATE requires an additional step of averaging predictions over the empirical distribution of the confounders \( X \). That is, we posit a model such as

\[
E[Y|Z, X; \beta] = \beta_X^T h(X, Z),
\]

where \( \beta_{X,Z} \) is a vector of parameters and \( h(X, Z) \) is a design matrix containing the main effects of the treatment, the covariates, and transformations of both, which could include interactions and polynomial functions. Assuming this model is structural (captures the true data-generating mechanism and so models the potential outcomes), the ATE may be estimated from a sample of size \( n \) as

\[
\hat{E}[Y(1) - Y(0)] = n^{-1} \sum_{i=1}^{n} \hat{Y}_i(1) - n^{-1} \sum_{i=1}^{n} \hat{Y}_i(0)
\]

\[
= n^{-1} \sum_{i=1}^{n} \hat{\beta}_{X,Z}^T h(X_i, 1) - n^{-1} \sum_{i=1}^{n} \hat{\beta}_{X,Z}^T h(X_i, 0).
\]

If the estimand of interest was the ATT rather than the ATE, the marginalization (empirical averaging) would simply be over only those individuals observed to be treated in the sample.
rather than the entire sample. Thus, the average would be taken over the distribution of $X | Z = 1$. There is no computational or conceptual barrier to computing a predicted outcome $\hat{\beta}^\top X$, that is under the condition $Z = 0$, for those individuals for whom $Z$ is in fact observed to be 1.

### 4.3.2. Propensity score regression

Propensity score regression can be viewed as a special case of outcome regression modelling, although the assumptions required for consistency differ. Under this approach, a model is specified as in Equation (1). However, in this specification, the design matrix $h(X, Z)$ includes as columns (“predictors”) the main effect of the treatment and the estimated propensity score: for example,

$$E[Y|Z, X, \beta] = \beta_0 + \beta_Z Z + \beta_e e(X). \quad (2)$$

If the treatment effect is modified by any covariates in $X$, then those interactions and an interaction between the treatment and the propensity score also must be included. The propensity score is typically estimated using logistic regression (Alam, Moodie & Stephens, 2019), with $\hat{e}(x)$ taken as fitted values from that model. Unlike traditional outcome regression modelling, we need not assume that Equation (1) is correctly specified, but rather, that the propensity score model is correctly specified and that any heterogeneity in the treatment effect is adequately captured by including interactions in $h(X, Z)$. Specifically, if

$$h(X, Z) = \beta_0^\top h_0(X) + \beta_1^\top h_1(X, Z),$$

then the treatment-free component $h_0(X)$ may be misspecified provided that the correct propensity score regression model is constructed. Suppose, for example, that the data-generating model has a conditional mean that contains an interaction with the confounder $X_1$:

$$E[Y|Z, X, \beta] = \beta_0 + \beta_{Z0} Z + \beta_{Z1} ZX_1.$$

The propensity score regression model must then contain terms that block the open backdoor paths that pass through the interaction; for example, this can be achieved using the model

$$E[Y|Z, X, \beta] = \beta_0 + \beta_{Z0} Z + \beta_{e0} e(X) + \beta_{Z1} ZX_1 + \beta_{e1} X_1 e(X).$$

We expand upon this issue and connections with semiparametric estimation in Section 5.

### 4.3.3. Propensity score stratification or matching

While propensity score regression can be viewed as a means of parametric smoothing to make comparisons between individuals with different values of the treatment who share the same value of the propensity score, the nonparametric counterpart seeks to create strata of individuals with similar covariates and estimate the treatment effect within these strata, ultimately averaging the effect over the strata. If $X$ is very low dimensional and is composed of discrete variables, it may be feasible to create distinct strata for each combination of the covariates without running into issues due to data sparsity. However, as the dimensionality of $X$ increases, this is not possible. The propensity score simplifies the analysis by providing a univariate summary of $X$ that can be used to stratify the sample. This may be done coarsely, such as by using quintiles of the propensity score distribution (Rosenbaum & Rubin, 1984), or finely using, say, individually matched pairs (Rubin, 1974).

If matching rather than stratifying, additional considerations must be taken into account (Stuart, 2010). First, a matching criterion must be selected, such as the nearest neighbour or a random draw from within some “caliper” distance of the propensity score. A computational
algorithm (greedy matching, optimal matching, etc.) must also be selected. Matching need not
be in pairs but could, for example, be 2:1 or 5:1. Matching is typically done without replacement,
which prioritizes bias over efficiency.

Finally, an important note on implementation: matching is so often used for estimating
the ATT that this is the default procedure in many software packages such as MatchIt and
Matching in R. Thus, when employing matching, one must take care to ensure that the method
used is appropriate for the chosen estimand.

4.3.4. Inverse probability of treatment weighting

As noted in Section 3.3, inverse weighting was first developed to address missing data and then
subsequently used to address confounding by viewing potential outcomes through a missing
data lens. After estimating the propensity score, inverse probability of treatment weighting
(IPTW) proceeds by fitting weighted averages (or a weighted regression model) using the
weights \( w_i = \frac{1}{\hat{e}(x_i)} + (1 - z_i) \left( 1 - \hat{e}(x_i) \right)^{-1} \). That is, each observation is weighted by the inverse
probability of having received the observed treatment (not the \( Z = 1 \) level of treatment). This
is often described as constructing a “pseudosample” in which there are no imbalances in the
distribution of the confounder between treatment groups (Hernán, Brumback & Robins, 2000).
IPTW can also be motivated by an importance sampling argument in which the target distribution
has \( Z \) independent of \( X \) (Saarela et al., 2015). Similar arguments can be extended to estimate
the ATT (Moodie, Saarela & Stephens, 2018).

While IPTW relies on the correct specification of the propensity score only, it may be very
inefficient relative to a correctly specified outcome regression model. The augmented IPTW
estimator reduces variability in the IPTW estimator and simultaneously provides additional
safeguards against model misspecification (Scharfstein, Rotnitzky & Robins, 1999; Bang &
Robins, 2005). An augmented estimator can also be computed by fitting an unweighted
regression model that includes the difference in the inverse probability of the treatments, i.e.,
\( \frac{\hat{e}(x_i)}{(1 - z_i)(1 - \hat{e}(x_i))^{-1}} \), as a covariate. See Section 4.4 for more on double robustness.

4.3.5. Model checks

All the methods described above rely on identifiability assumptions that include no unmeasured
confounding and positivity. They further make assumptions about correct model specification,
whether for the outcome model or the propensity score model. When the estimation approach
relies on correct specification of the outcome model, all standard residual diagnostics should be
employed, although it can be difficult to detect some forms of misspecification, and positivity
cannot be reliably assessed (Kang & Schafer, 2007).

When using propensity score methods, checking that balance between treatment groups has
been achieved can be accomplished by visually comparing distributions of covariates (e.g.,
Tan, 2006) or using distributional summaries such as the standardized mean difference (Higgins,
Li & Deeks, 2022). Positivity can be assessed through a visual examination of the propensity
score distribution by checking whether the distributions among the different treatment groups
overlap. For a worked example using all of the above-described methods that illustrates the
model checks, see the tutorial of Goetghebeur et al. (2020). There is no straightforward way to
assess positivity when \( X \) is even moderately high dimensional without relying on a propensity
score. Thus, while traditional methods of statistical analysis such as regression can be employed
for estimation, the propensity score still proves useful.

4.4. Double Robustness

The estimation procedures described above typically depend on the specification of two models
for the observable quantities: the outcome model, representing the conditional distribution

DOI: 10.1002/cjs.11718  The Canadian Journal of Statistics / La revue canadienne de statistique
of $Y$ given $(X, Z)$; and the treatment model, representing the conditional distribution of $Z$ given $X$. Doubly robust procedures provide consistent estimation of the ATE if at least one of these models is correctly specified and matches the data-generating structure (at least with respect to treatment and confounding variables). The augmentation strategy from Section 4.3.4 is an example of a doubly robust procedure, as the resulting estimator is consistent provided that at least one of the “augmenting” (outcome) model or the propensity score-based reweighting model based on the propensity score is correctly specified. Similarly, elaborations of the propensity score regression model in Equation (2) can also be considered doubly robust. If the outcome model

$$E[Y|Z, X, \beta] = \beta_0 + \beta_Z Z + \beta_e e(X) + \mu_0(X, \beta_X)$$

(3)

is used and, in reality, the data-generating model has an outcome such that the true conditional mean is $\beta_Z Z + \mu_0(X, \beta_X)$, then the least squares estimator for $\beta_Z$ derived from Equation (3) is also consistent even if the model for $e(x)$ is misspecified.

The lure of doubly robust procedures is evident. However, we must remember that any allusion to “correct specification of the outcome model” (one of the aforementioned aspects of robustness) is in conflict with the principal reason why propensity score adjustment is deployed. If we acknowledge the possibility that the outcome model is correctly specified, then by standard theory, the optimal approach to inference is the one that uses this outcome model without any further elaboration.

5. CAUSAL INFERENCE UNDER SEMIPARAMETRIC SPECIFICATIONS

Estimation using the models referred to in Section 4.3 is typically carried in a semiparametric inference setting without distributional assumptions about outcome residual errors or confounders. For linear regression models such as that in Equation (2), estimation is typically performed via least squares. Consequently, the optimality of approaches to inference in terms of efficiency is an important area of study. The form of optimal (if not always feasible) estimators of causal parameters can be deduced using semiparametric efficiency theory: see, for example, Bickel et al. (1998). This is now much studied and relied upon in the causal inference literature and has close connections with a rich history of techniques that were developed as extensions of parametric inference methods such as estimating functions in general (Godambe, 1991) and generalized estimating equations (Liang & Zeger, 1986) in particular, as well as parallel developments in econometrics (see Newey (1990) for a summary).

The model in Equation (2) provides a simple setting in which semiparametric results can be illustrated. Suppose that the causal (structural) model is

$$Y_i = \beta_Z Z_i + h_0(X_i) + \epsilon_i,$$

(4)

where $h_0(X_i)$ is some function of the confounders that is unknown to the analyst and $\epsilon_i$ satisfies $E[\epsilon_i|X_i] = 0$ and $\text{Var}[\epsilon_i|X_i] = \sigma^2(X_i) < \infty$. The key paper of Robins, Mark & Newey (1992) shows that the estimating equation

$$\sum_{i=1}^n (z_i - e(x_i))(y_i - \beta_Z z_i) = 0$$

(5)

with solution

$$\hat{\beta}_Z = \frac{\sum_{i=1}^n (z - e(x_i))y_i}{\sum_{i=1}^n (z - e(x_i))z_i}$$
results in consistent inference for $\beta_Z$; see also Newey (1990). The estimating equation in Equation (5) does not attempt to model the treatment-free component $h_0(X)$, but by considering the (inference) regression model

$$Y_i = \beta_Z z_i + \mu_0(X_i, \phi) + \epsilon_i,$$

(6)

the estimating equation can be modified to

$$\sum_{i=1}^{n} (z_i - e(x_i)) (y_i - \beta_Z z_i - \mu_0(x_i, \hat{\phi})) = 0,$$

yielding the solution

$$\hat{\beta}_Z = \frac{\sum_{i=1}^{n} (z_i - e(x_i)) (y_i - \mu_0(x_i, \hat{\phi}))}{\sum_{i=1}^{n} (z_i - e(x_i)) z_i},$$

which corresponds to an estimator that typically has a variance that is no larger than that of the original estimator.

There is a connection between this approach and the propensity score regression model in Equation (2). Suppose for simplicity that the intercept in Equation (2) is omitted and that the parameters are estimated using least squares. The corresponding estimating system becomes

$$\sum_{i=1}^{n} (z_i - e(x_i)) (y_i - \beta_Z z_i - \beta_e e(x_i)) = 0.$$

Subtracting the second equation from the first yields the equivalent form

$$\sum_{i=1}^{n} (z_i - e(x_i)) (y_i - \beta_Z z_i - \beta_e e(x_i)) = 0.$$

Thus, choosing $\mu_0(x, \phi) = \phi e(x)$ in Equation (6) corresponds to the model in Equation (2). Extension to more complicated linear specifications, including those that involve interactions between a treatment and confounders, is straightforward. We may conclude that the propensity score regression approach is a version of the semiparametric formulation.

Equation (5) is based on the $m$-estimating function

$$m(X, Y, Z; \beta_Z) = (Z - E[Z|X])(Y - \beta_Z Z)$$

and it is evident that, provided the term $(Y - \beta_Z Z)$ is independent of $Z$ (i.e., specifically, that the effect of treatment is correctly captured in Equation (4)), $E[m(X, Y, Z; \beta_Z)] = 0$ if the model $E[Z|X] = e(X)$ is correctly specified. In these calculations, expectations are taken with respect to the structural (data-generating) model.

As summarized by Newey (1990) (see also Tsiatis, 2006, Chapter 3), the semiparametric efficient estimator of $\beta_Z$ can be obtained by considering the influence function

$$\varphi(X, Y, Z; \beta_Z) = -\frac{1}{E[m(X, Y, Z; \beta_Z)]} m(X, Y, Z; \beta_Z),$$

where

$$\dot{m}(x, y, z; \beta_Z) = \frac{\partial m(x, y, z; \beta_Z)}{\partial \beta_Z} = -z(z - e(x)),$$
and then solving the estimating equation

\[ \sum_{i=1}^{n} \varphi(x_i, y_i, z_i; \beta_Z) = 0. \]

This equation matches Equation (5) as \( \hat{m}(x, y, z; \beta_Z) \) does not depend on \( \beta_Z \). Thus, under standard conditions, the estimator that results from Equation (5) is consistent and asymptotically normally distributed with asymptotic variance

\[ \frac{E[(m(X, Y, Z; \beta_Z))^2]}{E[\hat{m}(X, Y, Z; \beta_Z)]^2} = \frac{E[(Z - E[Z|X])^2(Y - \beta_Z Z)^2]}{E[Z - e(X)]^2} = \frac{E[\text{Var}[Z|X]\nu(X)]}{E[\text{Var}[Z|X]]}^2, \]

where \( \nu(X) = E[(Y - Z\beta_Z)^2|X] = \{h_0(X)\}^2 + \sigma^2(X) \). In this expression, \( \text{Var}[Z|X] = e(X)(1 - e(X)) \), and it is evident that, if \( \sigma^2(X) = \sigma^2 \) and there is a homoscedastic residual error structure, then

\[ \frac{E[\text{Var}[Z|X]\nu(X)]}{E[\text{Var}[Z|X]]}^2 \geq \frac{\sigma^2}{E[e(X)(1 - e(X))]} . \]

This provides a lower bound on the variance of the propensity score regression estimator.

A similar semiparametric efficiency analysis of the IPTW estimator from Section 4.3.4 can be carried out: see, for example, Hirano, Imbens & Ridder (2003), who provide the asymptotic variance of the IPTW estimator of the ATE \( \beta_Z \) in the model of Equation (4) as

\[ E \left[ \frac{\sigma^2_0(X)}{1 - e(X)} + \frac{\sigma^2(X)}{e(X)} \right] , \]

where \( \sigma^2_j(X) \) is the residual variance for \( Z = j \) for \( j = 0, 1 \). Under homoscedasticity assumptions, this reduces to

\[ \sigma^2 E \left[ \frac{1}{e(X)(1 - e(X))} \right] . \]

There are several things to note about these results. First, the lower bound for the regression estimator is smaller than the lower bound for the IPTW estimator, and if in fact \( h_0(X) \equiv 0 \), the lower bound is achieved. This result was noted by Ertefaie (2011). Second, the IPTW method does not rely on correct specification of the treatment effect term. Finally, these results for both methods can be extended to procedures involving augmentation and also to cases where the treatment effect model reflects modification by (i.e., interaction with) confounders that include additional treatment effect parameters.

If the propensity score model is represented by a parametric form, say \( e(x) \equiv e(x; \gamma) \), then in the estimation we may replace \( e(x) \) by \( \hat{e}(x) \equiv e(x; \hat{\gamma}) \), where \( \hat{\gamma} \) is obtained from an estimating procedure that depends on the \( x \) and \( z \) data only, and proceed with the “feasible” estimation that solves the corresponding version of Equation (5). As shown by Henmi & Eguchi (2004), the use of this plug-in estimator does not affect the asymptotic variance, as \( \hat{\beta}_Z \) and \( \hat{\gamma} \) are asymptotically independent.

6. MODERN DEVELOPMENTS AND CURRENT CHALLENGES

In this section, we point to some current directions for research in causal inference. In some instances, these directions involve weakening or violations of the identifiability assumptions listed above. Other research directions rely on the traditional identifiability assumptions and require additional but similar assumptions to further account for new data structures that have arisen in step with increases in digital storage capacity or computational power.
6.1. Unmeasured Confounding

The presence of unmeasured confounding disrupts causal (and indeed most conventional) statistical analyses. Unmeasured confounders are the “unknown unknowns” of statistics: we do not know if they exist, and furthermore, we cannot always be sure that we will ponder their existence. The left panel of Figure 3 demonstrates the simplest setting with unmeasured confounding; even if \(X\) is conditioned upon, the open path from \(Z\) to \(Y\) via \(U\) allows confounding of the treatment effect. In a standard analysis, there is no way of overcoming this unmeasured confounding as data on \(U\) are not available. It is possible to examine the effect of a possible hypothetical unmeasured confounder using simulation-based methods in a sensitivity analysis. An alternative approach can be derived if there exists a further (observable) variable \(W\) that (a) is a cause of \(Z\), (b) is independent of \(U\), and (c) has no direct effect on \(Y\): see the right panel of Figure 3.

For example, consider a randomized experimental study of a binary treatment with imperfect compliance. In such a study, there are no measured confounders. Study participants are randomly assigned to one of two treatment groups and each participant may adhere to or contravene the treatment assignment. If \(W\) is the treatment assignment indicator and \(Z\) records the treatment actually taken, then \(W\) is an instrument. Even if adherence to the assigned treatment is determined by an unmeasured factor \(U\), \(W\) is independent of \(U\) by design. Of course, an intention-to-treat analysis, which compares outcomes for the two assigned treatments, can still be carried out, but it may not reflect the target of inference.

Instrumental variable methods are sometimes claimed to be able to “overcome” the issue of unmeasured confounding. Perhaps a more honest view is that instrumental variable methods solve the causal inference problem by relying on a different set of equally strong assumptions from those listed in Section 4. In particular, it is only under strong assumptions that an instrumental variable analysis can be used to estimate an ATE or ATT. Otherwise, the analysis targets the complier average causal effect (CACE), which is the effect of treatment among those individuals whose observed treatment agrees with their assigned (by the instrument) treatment—a latent subgroup that complicates interpretation.

There are two connections with the methods described earlier that are noteworthy. First, the ATT in Section 3.4 may be estimated using instrumental variable methods: see Imbens & Angrist (1994) and Angrist, Imbens & Rubin (1996). Second, the propensity score regression estimator derived from Equation (5) can also be justified as an instrumental variable estimator using the instrument \(W = Z - e(X)\).

6.2. Interference

Interference occurs when an individual’s potential outcome depends on not only their treatment assignment but that of others as well. The presence of interference is a clear violation of SUTVA and requires an extended potential outcome notation and specialized methods. Currently, estimation has focused on methods of simplifying the way in which interference can occur, such that each individual’s treatment can be viewed as a bivariate vector whose components consist of
the treatment received directly by that individual and that received indirectly via spill-over. The latter could represent a proportion or the number of treated “others,” often termed neighbours, who affect that individual. Estimands focus on the impact of direct treatment assuming a fixed level of interference (e.g., the average outcome if treated vs not, given that 30% of one’s neighbours are treated), as well as the indirect impact of treatment (e.g., the average outcome if untreated and, say, 50% of one’s neighbours are treated as compared to if untreated and none of one’s neighbours is treated).

Hudgens and Halloran (Hudgens & Halloran, 2008; Halloran & Hudgens, 2012; Halloran & Hudgens, 2016; Saul, Hudgens & Halloran, 2017) have developed many critical innovations in this realm, including the notion of partial interference, which supposes that the data are composed of many (often small) fully connected clusters of individuals such that spill-over occurs within but not between clusters. Alternatively, some authors have relaxed SUTVA (van der Laan, 2014; Forastiere, Airoldi & Mealli, 2021) and instead rely on the stable unit treatment on neighbourhood value assumption (SUTNVA). SUTNVA assumes that, within a social network, interference or spill-over of a treatment can occur only from an immediate connection, termed an individual’s “neighbourhood.” This literature requires knowledge of the network of connections, which could be defined by, for example, self-reported friendships or social media connections.

6.3. Sequential Multiple Assignment Randomized Trials
An area of growing interest is that of algorithmic decision-making and decision-support systems. While these may be ubiquitous in many online activities, the use of individually tailored decisions in other fields such as the health sector calls for a high level of care and scrutiny. Within the statistical literature, this area is often termed “precision medicine,” and methodologies have centred on estimation and inference for dynamic treatment regimes or adaptive treatment strategies. Approaches to estimate optimal treatment strategies, particularly for sequences of treatments in which there may be delayed effects and interactions between treatments taken at different points in time, have grown predominantly from the causal inference literature. These approaches often rely on nonexperimental data and feature challenges such as time-dependent confounding and mediation. However, there remains an important role for randomized trials, using sequential multiple assignment randomized trials (SMARTs) (Collins, Murphy & Strecher, 2007; Oetting et al., 2011; Lei et al., 2012; Kidwell, 2014) to study sequences of treatments tailored to individual patient characteristics.

SMARTs are growing in number and yet the design is still not common and further innovations are needed. SMARTs are often used to determine optimal medication sequences or educational interventions that may best be delivered via a cluster randomized trial, e.g., enlisting physicians who will recruit patients from their clinic in the first case or randomizing educators (or schools) in the second. While a very small number of cluster SMARTs have been carried out (Kilbourne et al., 2014), they are rare and their design operating characteristics have not yet been fully explored. Further, adding adaptive components to the trial design may be of interest to accommodate early stopping (Cheung, Chakraborty & Davidson, 2015).

6.4. Electronic Health Records and Other Not-for-Research Data Sources
Data are collected and stored from an increasingly wide variety of sources. For example, many individuals voluntarily share data by using or uploading information to fitness tracking apps or through loyalty programs at retailers. While ever-increasing storage and computational capacity has created opportunities for new analyses, the increased availability and quantity of data does not necessarily correspond with an increase in its quality, so causal thinking—together with analyses aimed at reducing or eliminating biases—is required.

Electronic health records (EHRs) maintain health information on an often large population and log data on interactions with a healthcare provider, laboratory and other physiological
measurements, and prescribed treatments (Ehrenstein et al., 2019). These data may be subject to biases stemming from a variety of causes including lack of generalizability, covariate-driven patterns of observation (or missingness), measurement error and misclassification, and confounding. Further, statistical methods to ensure data security and privacy may need to be developed if EHRs from multiple sites are combined.

The generalizability of EHR data will typically depend on the source of the data. EHRs drawn from health registries in countries or regions with national or single-payer healthcare (e.g., the National Health Service in the United Kingdom or provincial healthcare in Canada) are more likely to be representative of the population under study than are registries of privately or employer-insured populations, or of those receiving subsidized medical care (e.g., Medicaid in the United States).

A common feature of EHRs, particularly with data on a diverse population, is that observations may be irregular. Some patients may be observed frequently because of a high number of comorbid conditions or medications that require close monitoring. Other patients may be observed less frequently, e.g., individuals with inflexible working hours. These covariate-driven monitoring patterns can lead to certain patient groups being over-represented in the data which, if not addressed in an analysis, can bias estimators (Coulombe, Moodie & Platt, 2021; Coulombe et al., 2022).

In terms of measurement error or misclassification, EHRs may be subject to misclassification of medication usage, as this is not often available in these records. Rather, EHRs typically record drugs prescribed, but might not record drugs dispensed (pharmacy claims). Where information on both the prescription and acquisition of a treatment is available, actual usage is rarely known and must be inferred from algorithms that make assumptions regarding overlapping prescriptions and gaps (grace periods) between prescriptions. For example, two treatments may be regarded as consecutive or maintained if no more than 4 weeks pass between the end of one prescription and the beginning of another, and otherwise be deemed separate treatment episodes (Macleod et al., 2019).

6.5. The Role of Machine Learning in Causal Inference

Machine learning methods are revolutionizing many applied fields and are at the centre of a huge research enterprise encompassing aspects of statistics, mathematics, and computer science. Statisticians play a vital role in this enterprise, particularly in developing frameworks for understanding the operating characteristics of machine learning methods when applied to data. Many aspects of the most successful machine learning approaches have their foundations in the statistical literature (e.g., hierarchical models, Gaussian processes, stochastic approximation methods, Dirichlet processes). The current applicability of these methods, at such scale and in such complexity, requires sophisticated algorithmic strategies as well as developments in theoretical understanding that have originated in the machine learning literature.

The impact of machine learning, and artificial intelligence in general, in health applications, for example, is already widely felt in the public and private sectors. These fields in fact have a long history, stemming at least from developments in the probabilistic and computational treatment of graphical models and Bayesian networks originating more than 40 years ago (see, e.g., Shortliffe & Buchanan, 1975). There are, however, limits to what can be achieved by the “automatic” application of mathematical, statistical, or computational approaches alone. In the current context, it is not wholly unfair to characterize many modern machine learning methods as prediction machines that can produce superlative reconstructions of observed data and considerable ability to generalize those out-of-sample predictions. Whereas this has evident attraction in many domains, it is also evident that prediction is not at the heart of causal reasoning.

From one perspective, the use of machine learning methods as flexible prediction tools can be helpful in constructing the models described in Section 4.3, which can be then deployed to estimate treatment effects. See Kennedy (2016) for a clear review of the empirical process

DOI: 10.1002/cjs.11718
theory that governs the behaviour of nonparametric estimators of nuisance functions in causal inference. However, the use of machine learning tools needs to be done with care, as the statistical properties of flexible fitting procedures are not always straightforward to understand.

Where to deploy machine learning within the causal framework also requires careful thought. Better estimation or prediction of the outcome model can lead to superior performance with reduced variability or bias in doubly robust or outcome modelling-based methods (see, e.g., Athey & Imbens, 2016; Hahn, Murray & Carvalho, 2020). However, uncertainty representation (and consequently testing) based on machine learning-based fitting needs very careful calibration. Kernel-based methodologies (Dai & Li, 2022) offer a way forward, as they are flexible yet statistically tractable. Further, using machine learning methods to construct the treatment assignment/propensity score model requires particular care and is arguably ill-advised because the primary objective of propensity score modelling is to create balance. A propensity score that predicts treatment with high precision may lead to a lack of overlap and hence positivity violations (Alam, Moodie & Stephens, 2019). Also, as in other domains, there are ethical aspects to consider when using machine learning-based analyses (Shortreed & Moodie, 2020): for example, algorithm-generated treatment decisions that are not transparent in their construction and may perpetuate errors made in the training data.

7. CONCLUDING REMARKS

As we have elaborated above, causal inference in statistics has formalized, mathematized, and operationalized notions of fairness so that causal conclusions can be drawn from analyses using imperfect data that are subject to confounding, missing data or selection biases, measurement error, and more. Causal inference should not be seen as conflicting with traditional statistical methods of explanatory analyses, and causal thinking should be considered relevant even in predictive modelling or machine learning applications. Not only can the underlying data-generating structure lead to potential biases such as confounding, but the data collection and research design can also give rise to inappropriate inference and decision making if not appropriately taken into account.

ACKNOWLEDGEMENTS

The authors acknowledge funding from Discovery Grants from the Natural Sciences and Engineering Research Council of Canada. EEMM is supported by a career award from the Fonds de recherche du Québec–Santé and holds a Canada Research Chair in Statistical Methods for Precision Medicine.

REFERENCES

Alam, S., Moodie, E. E. M., & Stephens, D. A. (2019). Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. Statistics in Medicine, 38(9), 1690–1702.
Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91, 444–455.
Athey, S. & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353–7360.
Bang, H. & Robins, J. M. (2005). Doubly-robust estimation in missing data and causal inference models. Biometrics, 61, 962–972.
Bickel, P., Klaassen, C., Ritov, Y., & Wellner, J. A. (1998). Efficient and Adaptive Estimation for Semiparametric Models, Springer, New York.
Bradford Hill, A. (1937). Principles of medical statistics. The Lancet. London, UK.
Bradford Hill, A. (1962). Alfred Watson Memorial Lecture: The statistician in medicine. Journal of the Institute of Actuaries, 88, 178–191.
Bradford Hill, A. (1965). The environment and disease: Association or causation? *Proceeding of the Royal Society of Medicine*, 58, 295–300.

Bradford Hill, A. (2020). The environment and disease: Association or causation? *Observational Studies*, 6, 1–65.

Bradford Hill, A. (2022). Alfred Watson Memorial Lecture: The statistician in medicine. *Journal of the Institute of Actuaries*, 21, 3–63.

Chalmers, I. (2011). Why the 1948 MRC trial of streptomycin used treatment allocation based on random numbers. *Journal of the Royal Society of Medicine*, 104(9), 383–386.

Cheung, Y. K., Chakraborty, B., & Davidson, K. W. (2015). Sequential multiple assignment randomized trial (SMART) with adaptive randomization for quality improvement in depression treatment program. *Biometrics*, 71, 450–459.

Cochran, W. G. & Rubin, D. B. (1974). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics Series A*, 35, 417–446.

Collins, L., Murphy, S. A., & Strecher, V. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): New methods for more potent e-health interventions. *American Journal of Preventive Medicine*, 32, S112–S118.

Coulombe, J., Moodie, E. E. M., & Platt, R. W. (2021). Weighted regression analysis to correct for informative monitoring times and confounders in longitudinal studies. *Biometrics*, 77, 162–174.

Coulombe, J., Moodie, E. E. M., Platt, R. W., & Renoux, C. (2022). Estimation of the marginal effect of antidepressants on body mass index under confounding and endogenous covariate-driven monitoring times. *Annals of Applied Statistics*, to appear.

Cox, D. R. (1958). *Planning of Experiments*, John Wiley and Sons, New York.

Dai, X. & Li, L. (2022). Orthogonalized kernel debiased machine learning for multimodal data analysis. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2021.2013851

Danaei, G., García Rodríguez, L. A., Cantero, O. F., Logan, R. W., & Hernán, M. A. (2018). Electronic medical records can be used to emulate target trials of sustained treatment strategies. *Journal of Clinical Epidemiology*, 96, 12–22.

Didelez, V. (2018). Causal concepts and graphical models. In Maathuis, M. H., Drton, M., Lauritzen, S. G., & Wainwright, M. J. (Eds.), *Handbook of Graphical Models*, Chapman & Hall/CRC Press, New York.

Ehrenstein, V., Kharrazi, H., Lehmann, H., & Taylor, C. O. (2019). Obtaining data from electronic health records. In Gliklich, R., Leavy, M., & Dreyer, N. (Eds.), *Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User’s Guide*, 3rd ed., Agency for Healthcare Research and Quality (US), Rockville.

Ertefaie, A. (2011). *Causal Inference via Propensity Score Regression and Length-Biased Sampling*. Ph.D. thesis, McGill University.

Fisher, R. A. (1951). *The Design of Experiments*, Oliver & Boyd, Edinburgh.

Fisher Box, J. (1980). R.A. Fisher and the design of experiments, 1922–1926. *The American Statistician*, 34(1), 1–7.

Forastiere, L., Airoldi, E. M., & Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534), 901–918.

Godambe, V. P. (1991). *Estimating Functions*, Oxford Science Publications, Clarendon Press, Oxford, UK.

Goetheheuer, E., le Cassie, S., Stavola, B. D., Moodie, E. E. M., & Waenbaum, I. (2020). Formulating causal questions and principled statistical answers. *Statistics in Medicine*, 39(30), 4922–4948.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10, 37–48.

Greenland, S. and Rafi, Z. (2020). To aid scientific inference, emphasize unconditional descriptions of statistics. arXiv preprint, arXiv:1909.08583.

Greenland, S. & Robins, J. (2009). Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations*, 6, 1–9.

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3), 965–1056.

Halloran, M. E. & Hudgens, M. G. (2012). Causal inference for vaccine effects on infectiousness. *The International Journal of Biostatistics*, 8(2), 1–40.

DOI: 10.1002/cjs.11718
Halloran, M. E. & Hudgens, M. G. (2016). Dependent happenings: A recent methodological review. Current Epidemiology Reports, 3(4), 297–305.

Henmi, M. & Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. Biometrika, 91, 929–941.

Hernán, M. A., Brumback, B., & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology, 11, 561–570.

Hernán, M. A. & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. American Journal of Epidemiology, 183, 758–764.

Higgins, J. P. T., Li, T., & Deeks, J. (2022). Choosing effect measures and computing estimates of effect. In Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., & Welch, V. (Eds.), Cochrane Handbook for Systematic Reviews of Interventions, Cochrane, Chichester, UK, 143–176.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4), 1161–1189.

Holland, P. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, 945–970.

Hudgens, M. G. & Halloran, M. E. (2008). Toward causal inference with interference. Journal of the American Statistical Association, 103(482), 832–842.

Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. Econometrica, 62, 467–475.

Kang, J. D. Y. & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science, 22, 523–539.

Kennedy, E. (2016). Semiparametric theory and empirical processes in causal inference. In He, H., Wu, P., & Chen, D.-G. (Eds.), Statistical Causal Inferences and Their Applications in Public Health Research, Springer, New York.

Kidwell, K. M. (2014). SMART designs in cancer research: Past, present, and future. Clinical Trials, 11, 445–456.

Kilbourne, A., Almirall, D., Eisenberg, D., Waxmonskey, J., Goodrich, D., Fortney, J., Kirchner, J. et al. (2014). Protocol: Adaptive Implementation of Effective Programs Trial (ADEPT): Cluster randomized SMART trial comparing a standard versus enhanced implementation strategy to improve outcomes of a mood disorders program. Implementation Science, 9, 132.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., & Murphy, S. (2012). A SMART design for building individualized treatment sequences. Annual Review of Clinical Psychology, 8, 21–48.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1), 13–22.

Macleod, J., Steer, C., Tilling, K., Cornish, R., Marsden, J., Millar, T., Strang, J., & Hickman, M. (2019). Prescription of benzodiazepines, z-drugs, and gabapentinoids and mortality risk in people receiving opioid agonist treatment: Observational study based on the UK Clinical Practice Research Datalink and Office for National Statistics death records. PLoS Medicine, 16(11), e1002965.

Medical Research Council. (1949). Streptomycin treatment of pulmonary tuberculosis: A Medical Research Council investigation. British Medical Journal, 2, 769–782.

Moodie, E. E. M., Saarela, O., & Stephens, D. A. (2018). A doubly robust weighting estimator of the average treatment effect on the treated. Stat, 7, e205.

Moodie, E. E. M. & Stephens, D. A. (2010). Marginal structural models: Unbiased estimation for longitudinal studies. International Journal of Public Health, 56, 117–119.

Murray, E. J., Marshall, B. D., & Buchanan, A. L. (2021). Emulating target trials to improve causal inference from agent-based models. American Journal of Epidemiology, 190(8), 1652–1658.

Newey, W. K. (1990). Semiparametric efficiency bounds. Journal of Applied Econometrics, 5, 99–135.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay in principles. Section 9 (translation published in 1990). Statistical Science, 5, 465–472.

Oetting, A., Levy, J., Weiss, R., & Murphy, S. A. (2011). Statistical methodology for a smart design in the development of adaptive treatment strategies. In Shrout, P., Keyes, K., & Ornstein, K. (Eds.), Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures, American Psychiatric Publishing, Inc, Arlington, 179–205.

Pearl, J. (1995). Causal diagrams for empirical research. Biometrika, 82, 669–710.

Pearl, J. (2009). Causality, 2nd ed., Cambridge University Press, Cambridge.
Pearl, J. (2010). An introduction to causal inference. *International Journal of Biostatistics*, 6, 7.

Petrarca, F. (1364). *Rerum Senilium Libri*. Liber XIV, Epistola 1. (Letter to Boccaccio, V.3).

Robins, J. M. (1998). Correcting for non-compliance in randomized trials using structural nested mean models. In *1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, 1–10.

Robins, J. M. (1999a). Association, causation, and marginal structural models. *Synthese*, 121, 151–179.

Robins, J. M. (1999b). Marginal structural models versus structural nested models as tools for causal inference. In Halloran, M. E. & Berry, D. (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, Vol. 116, IMA, Springer, New York, 95–134.

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550–560.

Robins, J. M., Mark, S. D., & Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48, 479–495.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106–121.

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

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516–524.

Rubin, D. B. (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. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1), 1–26.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74, 318–328.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test (comment). *Journal of the American Statistical Association*, 75(371), 591–593.

Rubin, D. B. (1990). Comment on “On the application of probability theory to agricultural experiments”. *Statistical Science*, 5, 472–480.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20–36.

Saarela, O., Stephens, D. A., Moodie, E. E. M., & Klein, M. B. (2015). On Bayesian estimation of marginal structural models (with response to discussion). *Biometrics*, 71, 279–288.

Saul, B. C., Hudgens, M. G., & Halloran, M. E. (2017). Causal inference in the study of infectious disease. In Srinivasa Rao, A. S. R., Pyne, S., & Rao, C. R. (Eds.), *Disease Modelling and Public Health, Part A, Handbook of Statistics*, Volume 36. Elsevier, Amsterdam, Netherlands, 229–246.

Scharfstein, D., Rotnitzky, A., & Robins, J. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response model. *Journal of the American Statistical Association*, 94, 1096–1120.

Shimoni, Y., Karavani, E., Ravid, S., Bak, P., Ng, T. H., Alford, S. H., Meade, D., & Goldschmidt, Y. (2019). An evaluation toolkit to guide model selection and cohort definition in causal inference. arXiv preprint, arXiv:1906.00442.

Shortliffe, E. H. & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3), 351–379.

Shortreed, S. M. & Moodie, E. E. M. (2020). Automated analyses: Because we can, does it mean we should? *Statistical Science*, 35, 499–502.

Stuart, E. (2010). Matching methods for causal inference: A review and look forward. *Statistical Science*, 25, 1–21.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101, 1619–1637.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*, Springer, New York.

Upadhyayula, S. & Kasliwal, R. R. (2020). Commemorating James Lind: The man who gave us the concept of controlled clinical trials. *Journal of Clinical and Preventive Cardiology*, 9, 92–93.

DOI: 10.1002/cjs.11718
van der Laan, M. J. (2014). Causal inference for a population of causally connected units. *Journal of Causal Inference, 2*(1), 13–74.

VanderWeele, T. J., Hernán, M. A., & Robins, J. M. (2008). Causal directed acyclic graphs and the direction of unmeasured confounding bias. *Epidemiology, 19*, 720–728.

Received 4 February 2022
Accepted 22 March 2022