Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods

Yixin Fang

1 Introduction

In the opening article [1] of *Journal of Comparative Effectiveness Research*, the journal's founding editors pointed out that comparative effectiveness research (CER) "draws from the disciplines of health technology assessment, outcomes research, clinical epidemiology and implementation science, among others, to better answer the fundamental question 'which treatment will work best, in which patient, and under what circumstances?'"

Besides traditional randomized controlled clinical trials (RCTs), CER is looking at alternative real-world study designs [2], including:

- Pragmatic clinical trials such as pragmatic RCTs and large simple trials
- Observational studies such as case–control studies and cohort studies
- Non-randomized single-arm trials with external controls

In CER, causal inference plays an important role in deriving real-world evidence (RWE) from the analysis of real-world data (RWD) that are generated from real-world studies [3]. Research in causality has a long history, but in modern time, different disciplines (e.g., social science, economics, and statistics) took different paths. In this section, we provide a brief history of the development of causal inference in statistics before we move on to recent developments.

In *The Book of Why* [4], Pearl shared his regret that even the founding fathers of modern statistics such as Pearson hindered the development of causal inference in the community of statistics at the early stage of modern statistics. Since Neyman proposed the concept of potential outcomes in his 1923 Master's thesis and

Y. Fang (✉)
Data and Statistical Sciences, AbbVie, North Chicago, IL, USA
e-mail: yixin.fang@abbvie.com

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_11

Rubin in 1974 extended it into a general framework for causal inference in both interventional studies and non-interventional settings [5], we have seen more and more developments of causal-inference methods in the community of statistics. Counterfactual causal inference is the first one on the list of eight most important statistical ideas of the past 50 years selected by a 2021 paper [6]. Here we briefly review three milestones.

The first milestone is propensity-score (PS)-based methods developed by Rubin and colleagues, based upon a fundamental theorem proved in their 1983 paper [7]. The class of PS-based methods includes four methods: (1) matching, (2) stratification, (3) PS as covariate, and (4) weighting. The second milestone is generalized methods (G-methods) developed by Robins and colleagues in 1990s and 2000s, including three major methods: (i) g-formula, (ii) inverse probability of treatment weighting (IPTW), and (iii) G-estimation. Refer to their book [8] for a comprehensive review of G-methods. The third milestone is targeted learning developed by van der Laan and colleagues, starting with their first paper on targeted maximum likelihood estimation [9], leading to two books on targeted learning [10, 11].

The remaining of the chapter is organized as follows. There is a rich literature on reviewing and tutorials of causal inference methods, so we believe we cannot do better in providing another comprehensive review. Instead, in Sects. 2–4, we review some influential methods by making three binary choices: (a) conditional or marginal, (b) weighting or standardization, and (c) time-independent or time-dependent. In Sect. 5, we provide some discussion on the application of these methods to real-world studies with intercurrent events.

2 Conditional or Marginal

2.1 Propensity-Score Methods

We start with a simple point-exposure study, in which A is a binary exposure variable with $A = 1$ being the investigative treatment and $A = 0$ being the comparator (say, the standard of care), Y is an outcome variable that is either continuous or binary, and W is a list of covariates, which are believed to contain all the measured confounders along with effect modifiers. A directed acyclic graph (DAG) for this study is displayed in Fig. 1.

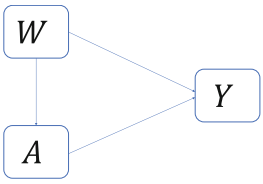

Fig. 1 A directed acyclic graph of a point-exposure study

We conduct causal inference to test the existence and estimate the magnitude of the relationship $A \rightarrow Y$, which is confounded by one back-door path [12], $A \leftarrow W \rightarrow Y$. The randomization feature in RCTs removes the arrow in $W \rightarrow A$, such that

$$A \perp\!\!\!\perp W, \tag{1}$$

leading to removing the confounding bias in the design stage. In non-randomized real-world studies, thanks to the following theorem in [7], we are able to achieve the desirable independence between A and W conditional on the PS function, $e(w) = P(A = 1|W = w)$.

Theorem 1 (Theorem 1 in [7]) *Treatment assignment and the observed covariates are conditionally independent given the propensity score, that is,*

$$A \perp\!\!\!\perp W|e(W). \tag{2}$$

There are four different PS methods based on the above theorem [13]: (1) matching on the PS, (2) stratification on the PS, (3) covariate adjustment using the PS, and (4) IPTW using the PS. Although the validity of all these four methods depends on whether or not PS function $e(w)$ is estimated consistently, in order to understand the pros and cons among them, it is helpful to understand the "conditional" thinking behind PS methods (1)–(3) and the "marginal" thinking behind PS method (4).

The first method, matching on the PS, attempts to mimic an RCT, creating a matched subset conditional on which A and W are independent. The second method, stratification on the PS, stratifies the dataset into several subsets, such that conditional on each subset, A and W are approximately independent. The third method, covariate adjustment using the PS, specifies a regression model of Y against A and $e(W)$, modeling the conditional relationship between Y and A given $e(W)$.

Unlike the first three PS methods that take the conditional thinking, IPTW takes the marginal thinking, creating two pseudo-populations, with one pseudo-population in which all the subjects were treated by $A = 1$ and the other pseudo-population in which all the subjects were treated by $A = 0$. Furthermore, of these four PS methods, IPTW is the only one that can be generalized to methods that can adjust for time-dependent confounding. Hence, we can consider IPTW as the intersection of the class of PS methods and the class of G-methods. IPTW is often discussed with marginal structural models (MSMs) [14], where we use MSMs to define an estimand and use the IPTW method to estimate the estimand.

2.2 Marginal Structural Models

Continue the above point-exposure study. Let $Y^{a=1}$ denote a subject's outcome if treated by the investigative treatment and $Y^{a=0}$ denote the outcome if treated by the comparator. For continuous outcome or 0–1 binary outcome, we can consider the following marginal structural models [14]:

$$E(Y^a) = \alpha + \beta a, \tag{3}$$

which are marginal models because they model the marginal distributions of potential outcomes $Y^{a=1}$ and $Y^{a=0}$ rather than the joint distribution, are structural models because they model the potential outcomes rather than the observed outcomes, and are saturated models because two unknown quantities ($E(Y^1)$ and $E(Y^0)$) are modeled by two parameters (α and β). Note that $\beta = E(Y^1) - E(Y^0)$ for continuous outcome or $\beta = P(Y^1 = 1) - P(Y^0 = 1)$ for binary outcome is the average treatment effect (ATE). In addition, for binary outcome, we may consider different MSMs, for example, $\text{logit}P(Y^a = 1) = \alpha' + \beta' a$, where β' is the log odds ratio between $Y^1 = 1$ and $Y^0 = 1$. Overall, the parameters in these MSMs can be estimated using the IPTW estimators [14].

Because of potential confounding, linear regression analysis of $Y \sim A$ for continuous outcome is biased in estimating β, and logistic regression analysis of $Y \sim A$ for binary outcome is biased in estimating β'. On the other hand, assuming that there is no unmeasured confounding, using weight $\omega = A/e(W) + (1-A)/(1-e(W))$, weighted linear regression analysis and weighted logistic regression analysis are unbiased in estimating β and β', respectively.

The approach of MSM and IPTW can be generalized to analyze studies with multi-level treatment, studies with continuous treatment doses, and studies with time-dependent confounding [14].

3 Weighting or Standardization

There is a rich literature of causal inference methods beyond the PS methods, which are well reviewed in several monographs (e.g., [8, 10, 11, 15, 16]). It is not our intention to review these recent developments comprehensively. Instead, as in [17], in this section, we describe two basic strategies, the weighting strategy and the standardization strategy.

We continue the above point-exposure study, which generates a dataset consisting of $O_i = (W_i, A_i, Y_i)$, $i = 1, \cdots, n$. In (3), the causal quantity is defined as parameter β in the MSM. Here we define the causal quantity of interest as the following ATE directly:

$$\theta^* = E(Y^1) - E(Y^0). \tag{4}$$

In order to construct an estimand, we assume three assumptions [8]: the consistency assumption, the no-unmeasured-confounder (NUC) assumption, and the positivity assumption,

$$\begin{aligned}\text{Consistency}: &\ Y = AY^1 + (1-A)Y^0,\\ \text{NUC}: &\ Y^a \perp\!\!\!\perp A|W, a = 0, 1,\\ \text{Positivity}: &\ P(A = a|W = w) > 0, a = 0, 1; w \in supp(W).\end{aligned}$$

In addition, we may need either or both of the following two functions, the PS function from the propensity-score model of $A \sim W$,

$$g(a|w) = P(A = a|W = w), \tag{5}$$

and the regression function from the outcome-regression model of $Y \sim A + W$,

$$Q(a, w) = E(Y|A = a, W = w). \tag{6}$$

3.1 The Weighting Strategy

3.1.1 Estimand

Under those three identifiability assumptions, we have

$$\begin{aligned}
&E\left\{\frac{I(A=a)}{P(A=a|W)}Y\right\} \quad \because \text{the positivity assumption}\\
&= E\left[E\left\{\frac{I(A=a)}{P(A=a|W)}Y\middle|W\right\}\right] \quad \text{by the double expectation formula}\\
&= E\left[E\left\{\frac{I(A=a)}{P(A=a|W)}Y^a\middle|W\right\}\right] \quad \because \text{the consistency assumption}\\
&= E\left[E\left\{\frac{I(A=a)}{P(A=a|W)}\middle|W\right\}E\{Y^a\middle|W\}\right] \quad \because \text{the NUC assumption}\\
&= E\left[E\{Y^a|W\}\right] \quad \because E(I(A=a|W)=P(A=a|W)\\
&= E(Y^a). \quad \text{by the double expectation formula}
\end{aligned}$$

Hence, we have

$$\theta^* = E(Y^1) - E(Y^0) = E\left\{\frac{I(A=1)}{P(A=1|W)}Y\right\} - E\left\{\frac{I(A=0)}{P(A=0|W)}Y\right\}. \tag{7}$$

This leads to the following estimand,

$$\theta = E\left\{\frac{I(A=1)}{g(1|W)}Y\right\} - E\left\{\frac{I(A=0)}{g(0|W)}Y\right\}. \tag{8}$$

We call this strategy of defining estimand as the weighting strategy because it uses the inverse of $g(a|w) = P(A = a|W = w)$ as the weights in the definition of the estimand. Using these weights, it creates two pseudo-populations: one pseudo-population in which all the subjects would have been treated by $a = 1$, leading to the first term in the right-hand side of (8), and the other pseudo-population in which all the subjects would have been treated by $a = 0$, leading to the second term.

3.1.2 Initial Estimator

If we obtain an estimator of the PS function, $\widehat{g}(a|w)$, using some statistical model, say logistic regression model, then we can obtain an initial estimator of θ, the IPTW estimator,

$$\widehat{\theta}_{IPTW} = \frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=1)}{\widehat{g}(1|W_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=0)}{\widehat{g}(0|W_i)}Y_i. \tag{9}$$

3.1.3 Doubly Robust Estimator

Although initial estimator $\widehat{\theta}_{IPTW}$ is asymptotically consistent if the model of $A \sim W$ is correctly specified in the construction of $\widehat{g}(a|w)$, it is not asymptotically efficient. Therefore, it is desirable to develop an augmented estimator that is asymptotically efficient under some model specification requirements.

According to semi-parametric efficiency theory (e.g., [10, 18]), the efficient score of estimating θ is given by

$$D(\theta; g, Q) = \frac{2A-1}{g(A|W)}[Y - Q(A, W)] + Q(1, W) - Q(0, W) - \theta. \tag{10}$$

Based on this efficient score function, we can apply the estimating equation approach to obtain an augmented estimator of θ, $\widehat{\theta}_{AIPTW}$, such that

$$\sum_{i=1}^{n} D(\widehat{\theta}_{AIPTW}; \widehat{g}, \widehat{Q})(W_i, A_i, Y_i) = 0, \tag{11}$$

where estimators $\widehat{g}$ and $\widehat{Q}$ are obtained by specifying some models of $A \sim W$ and $Y \sim A + W$, respectively.

Thus, by solving the estimating equation (11), we obtain the following augmented inverse probability of treatment (AIPTW) estimator [19]:

$$\widehat{\theta}_{AIPTW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(A_i=1)}{\widehat{g}(1|W_i)}Y_i - \frac{I(A_i=1)-\widehat{g}(1|W_i)}{\widehat{g}(1|W_i)}\widehat{Q}(1,W_i)\right)$$
$$-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(A_i=0)}{\widehat{g}(0|W_i)}Y_i - \frac{I(A_i=0)-\widehat{g}(0|W_i)}{\widehat{g}(0|W_i)}\widehat{Q}(0,W_i)\right). \quad (12)$$

According to the theory of estimating equations [19], $\widehat{\theta}_{AIPTW}$ is a doubly robust estimator; that is, it is asymptotically consistent if either the propensity-score model or the outcome-regression model is correctly specified, and it is asymptotically efficient if both models are correctly specified.

3.2 The Standardization Strategy

3.2.1 Estimand

Under those three identifiability assumptions, we have

$$\begin{aligned}
&E(Y^a)\\
&= E\{E(Y^a|W)\} \quad \text{by the double expectation formula}\\
&= E\{E(Y^a|A=a,W)\} \quad \because \text{the NUC assumption and positivity assumption}\\
&= E\{E(Y|A=a,W)\}. \quad \because \text{the consistency assumption}
\end{aligned}$$

Hence, we have

$$\theta^* = E(Y^1) - E(Y^0) = E_W\{E(Y|A=1,W) - E(Y|A=0,W)\}. \quad (13)$$

This leads to the following estimand:

$$\theta = E_W\{Q(1,W) - Q(0,W)\} = \int [Q(1,w) - Q(0,w)]dP_W(w), \quad (14)$$

where $P_W(w)$ is the probability distribution of W in the study population.

We call this strategy of defining estimand as the standardization strategy because it uses the standardization expectation over the marginal distribution of W of the study population, $E_W\{Q(a,W)\}$, for $a = 0, 1$.

3.2.2 Initial Estimator

If we obtain an estimator of the regression function, $\widehat{Q}(a, w)$, using some regression model, say generalized linear model, then we can obtain an initial estimator of θ,

$$\widehat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}[\widehat{Q}(1, W_i) - \widehat{Q}(0, W_i)] = \int [\widehat{Q}(1, w) - \widehat{Q}(0, w)]d\widehat{P}_W(w), \quad (15)$$

where $\widehat{P}_W$ is the empirical distribution of W, which is a non-parametric maximum likelihood estimator of P_W. Following [10], we call the above estimator as maximum likelihood estimator (MLE). To understand this, let $\theta = \theta(Q, P_W)$. If $\widehat{Q}$ and $\widehat{P}_W$ are MLEs of Q and P_W, respectively, then $\widehat{\theta}_{MLE} = \theta(\widehat{Q}, \widehat{P}_W)$ is MLE of $\theta = \theta(Q, P_W)$.

3.2.3 Doubly Robust Estimator

Although initial estimator $\widehat{\theta}_{MLE}$ is asymptotically consistent if the model of $Y \sim A + W$ is correctly specified in the construction of $\widehat{Q}(a, w)$, it may not be asymptotically efficient. Therefore, it is desirable to develop a targeted estimator that is asymptotically efficient.

The efficient score of estimating $\theta = \theta(Q, P_W)$ in (10) can be written as $D(Q, P_W, g)(W, A, Y)$, which equals

$$\frac{2A - 1}{g(A|W)}[Y - Q(A, W)] + Q(1, W) - Q(0, W) - \theta(Q, P_W). \quad (16)$$

Based on this efficient score function, [9] develops the targeted learning technique to obtained estimators $(\widehat{Q}^*, \widehat{P}^*_W, \widehat{g}^*)$ such that

$$\sum_{i=1}^{n} D(\widehat{Q}^*, \widehat{P}^*_W, \widehat{g}^*)(W_i, A_i, Y_i) = 0, \quad (17)$$

where $\widehat{P}^*_W = \widehat{P}_W$, the empirical estimator of P_W, and $\widehat{g}^*$ and $\widehat{Q}^*$ are some updated estimators of initial estimators $\widehat{g}$ and $\widehat{Q}$, respectively. Thus, we can construct the targeted maximum likelihood estimator (TMLE),

$$\widehat{\theta}_{TMLE} = \theta(\widehat{Q}^*, \widehat{P}_W) = \int [\widehat{Q}^*(1, w) - \widehat{Q}^*(0, w)]d\widehat{P}_W(w). \quad (18)$$

3.3 Implementation and Comparison

Consider the implementation of the aforementioned four estimators: IPTW, AIPTW, MLE, and TMLE. We can use SAS procedure "CAUSALTRT" to implement $\widehat{\theta}_{IPTW}$, $\widehat{\theta}_{MLE}$, and $\widehat{\theta}_{AIPTW}$, along with their statistical inferences. Please see the following skeleton of the SAS procedure:

```
PROC CAUSALTRT;
MODEL outcome = covariate_1 covariate_2 ... ;
PSMODEL treatment = covariate_1 covariate_2 ...;
RUN;
```

In the above SAS procedure, there are "PSMODEL" and "MODEL" statements: (1) if only a generalized linear model (GLM) of $A \sim W$ is specified in the "PSMODEL" statement, it implements $\widehat{\theta}_{IPTW}$, (2) if only a GLM of $Y \sim A + W$ is specified in the "MODEL" statement, it implements $\widehat{\theta}_{MLE}$, and (3) if two GLM models are specified in the "PSMODEL" and "MODEL" statements, respectively, it implements $\widehat{\theta}_{AIPTW}$.

Furthermore, we can consider flexible models other than GLM (say, super learner [20]) to obtain initial estimators $\widehat{Q}$ and $\widehat{g}$ to improve the chance of consistency in estimating functions Q and g. For this aim, we can use R function "tmle" in R package "tmle" [21] to implement $\widehat{\theta}_{TMLE}$, along with its standard error for conducting statistical inference. Please see the following skeleton of the R function:

```
tmle(Y, A, W,
  Q.SL.library = c("SL.glm", "tmle.SL.dbarts2", "SL.glmnet"),
  g.SL.library = c("SL.glm", "tmle.SL.dbarts.k.5", "SL.gam"),
  family = "gaussian", ...)
```

In the above R function, we see that we adopt the same set of notations for variable names and function names in this chapter (e.g., Y, A, W, g, Q) from the R package "tmle," which makes it easy for us to plug in values into the arguments. For example, the "Q.SL.library" argument allows us to specify a flexible super learner model for the Q function, with a default library consisting of generalized linear model (glm), discrete Bayesian additive regression tree (dbart), and glm model regularized by elastic net (glmnet), while the "g.SL.library" argument allows us to specify a super learner model for the g function, with a default library consisting of glm, dbart, and generalized additive model (gam). Besides there default options, we can prespecify other options for the super learner libraries, including highly adaptive lasso. In addition, the "family" argument can take on default value "gaussian" for continuous outcome and other value "binomial" for binary outcome.

Chapter 6 of [10] provides both theoretical comparisons and numerical comparisons (extensive simulations and case studies) between these four methods. Here we only summarize some comparisons briefly. First, AIPTW and TMLE are doubly robust versions of IPTW and MLE, respectively. Second, AIPTW relies on parametric modeling of Q and g, while TMLE allows for flexible modeling of Q and g using super learner. Third, MLE and TMLE are plug-in estimators, which

are more stable than the weighted estimators. Fourth, all the four methods are G-methods, which can be generalized to analyze longitudinal data with time-dependent confounding.

4 Time-Independent or Time-Dependent

In the above point-exposure study, the treatment status is determined at a single time (time zero) for all the subjects and the treatment effect does not need to make references to the time at which treatment occurs [8]. On the other hand, in longitudinal studies with time-dependent treatments or intercurrent events, we need to incorporate time explicitly [8].

Chapter "Personalized Medicine with Advanced Analytics" of this book will review statistical methods for personalized medicine and dynamic treatment regimes. In this chapter, we focus on longitudinal studies with static treatment regimes and intercurrent events.

Assume that there is one longitudinal study starting with baseline $t = 0$, along with follow-up visits, $t = 1, \cdots, T$. Assume that the primary endpoint Y is the outcome variable at the final visit T. Let $\overline{A} = (A_0, \cdots, A_{T-1})$ be the actually received treatment sequence and $\overline{A}_t = (A_0, \cdots, A_t)$ be the treatment up to t, $t = 0, \cdots, T-1$. Let W_0 be baseline covariates, W_t be the vector including time-dependent covariates and intermediate outcome, and $\overline{W}_t = (W_0, \cdots, W_t)$ be the vector consisting of all the observed history up to time t including baseline covariates, time-dependent covariates, and intermediate outcomes.

Let $\overline{a} = (a_0, \cdots, a_{T-1})$ be a given static treatment regime. At each time t, $a_t = 1$ stands for treated by the investigative treatment, 0 for the comparator, `NA` for treatment discontinuation, and 2 for some rescue medication. Two examples are $\overline{a} = \overline{1} = \text{rep}(1, T)$, which means the subject is initially treated by $a_0 = 1$ and throughout, and $\overline{a} = \overline{0} = \text{rep}(0, T)$, which means the subject is initially treated by $a_0 = 0$ and throughout.

Let $Y^{\overline{a}^0}$ be the potential outcome if the subject follows the static treatment regime $\overline{a}^0 = (a_0^0, \cdots, a_{T-1}^0)$. The population summary of $Y^{\overline{a}^0}$ is referred to as the value of $\overline{a}^0$ in [16]. For continuous or binary outcome variable, we define the value of $\overline{a}^0$ as

$$v^*(\overline{a}^0) = E\{Y^{\overline{a}^0}\}. \tag{19}$$

In order to construct an estimand for the evaluation of the value, $v^*(\overline{a}^0)$, we also need three identifiability assumptions [8], the consistency assumption,

$$Y^{\overline{a}^0} = Y \text{ if } \overline{A} = \overline{a}^0, \tag{20}$$

the static sequential exchangeability assumption (a.k.a., the NUC assumption),

$$Y^{\overline{a}^0} \perp\!\!\!\perp A_0 | W_0,$$
$$Y^{\overline{a}^0} \perp\!\!\!\perp A_t | (\overline{A}_{t-1}, \overline{W}_t), \text{ for } t = 1, \cdots, T-1, \tag{21}$$

and the positivity assumption,

$$P\left(\overline{A} = \overline{a}^0 | W_0 = w_0\right) > 0, \text{ for } w_0 \in supp(W_0). \tag{22}$$

Consider a longitudinal study that generates a dataset consisting of $O_i = (W_{0i}, A_{0i}, \cdots, W_{T-1,i}, A_{T-1,i}, Y_i), i = 1, \cdots, n$. In the following two subsections, we will describe four major estimators, IPTW, AIPTW, MLE, and TMLE, that are respectively generalized from those four G-estimators described in Sect. 3. For this aim, we define two series of functions.

Propensity-Score Modeling
Let $H_0 = W_0$ and $H_t = (\overline{W}_t, \overline{A}_{t-1})$ be the history up to t before making decision $A_t, t = 1, \cdots, T-1$. Define the PS functions from modeling $A_t \sim H_t$,

$$g_t(a|h_t) = P(A_t = a | H_t = h_t), t = 0, \ldots, T-1. \tag{23}$$

We can obtain an estimator of $g_t(a|h_t)$, $\widehat{g}_t(a|h_t)$, using some statistical model such as logistic regression model.

Outcome-Regression Modeling
We attempt to define regression functions from modeling $Y \sim A_t + H_t$, $t = 0, \ldots, T-1$. However, the outcome variable Y is measured after the final decision point $T-1$, which depends on decisions made between $t+1$ and $T-1$. Therefore, we should apply some special approach to define them. The most popular approach is the backward induction approach [16], which defines regression functions recursively from decision point $T-1$ to decision point 0.

At decision point $T-1$, define

$$Q_{T-1}(H_{T-1}, A_{T-1}) = E(Y | H_{T-1}, A_{T-1}), \tag{24}$$

which can be estimated using some regression model such as GLM, with its estimator denoted as $\widehat{Q}_{T-1}(h_{T-1}, a_{T-1})$. Note that $(h_{T-1}, a_{T-1}) = (\overline{w}_{T-1}, \overline{a}_{T-1})$. Next, define $\widetilde{Q}_{T-1}(H_{T-1}) = Q_{T-1}(H_{T-1}, a^0_{T-1})$, which is the expected outcome if the treatment at $T-1$ is consistent with the static treatment regime $\overline{a}^0$ at $T-1$ and which can be used as the model outcome variable at decision point $T-2$.

At decision point $t = T-2, \ldots, 1$, define

$$Q_t(H_t, A_t) = E(\widetilde{Q}_{t+1}(H_{t+1}) | H_t, A_t), \tag{25}$$

which can be estimated using some regression model such as GLM, with its estimator denoted as $\widehat{Q}_t(h_t, a_t)$. Note that $(h_t, a_t) = (\overline{w}_t, \overline{a}_t)$. Next, define

$\widetilde{Q}_t(H_t) = Q_t(H_t, a_t^0)$, which is the expected outcome if the treatments at decision points from t to $T-1$ are consistent with the static treatment regime $\overline{a}^0$ at decision points from t to $T-1$.

Finally, at decision point $t = 0$, define

$$Q_0(W_0, A_0) = E(\widetilde{Q}_1(H_1)|W_0, A_0), \tag{26}$$

which can be estimated using some regression model such as GLM, with its estimator denoted as $\widehat{Q}_0(w_0, a_0)$. Define $\widetilde{Q}_0(W_0) = Q_0(W_0, a_0^0)$, which is the expected outcome if the subject takes the static treatment regime $\overline{a}^0$ at all decision points from 0 to $T-1$.

4.1 The Weighting Strategy

4.1.1 Estimand

By the weighting strategy, we can define the corresponding estimand for the value of $\overline{a}^0$. That is, under those three identifiability assumptions, $\nu^*(\overline{a}^0)$ is equal to

$$\nu(\overline{a}^0) = E\left\{\frac{I[\overline{A} = \overline{a}^0]Y}{g_0(a_0^0|W_0)\prod_{t=1}^{T-1} g_t(a_t^0|\overline{W}_t, \overline{A}_{t-1})}\right\}, \tag{27}$$

where propensity-score functions g's are defined in (23).

4.1.2 Initial Estimator

If we obtain estimators of propensity-score functions, $\widehat{g}_t$, $t = 0, \cdots, T-1$, then we can obtain an initial estimator of $\nu(\overline{d})$,

$$\widehat{\nu}_{IPTW}(\overline{a}^0) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I[\overline{A}_i = \overline{a}^0]Y_i}{\widehat{g}_0(a_0^0|W_{0i})\prod_{t=1}^{T-1}\widehat{g}_t(a_t^0|\overline{W}_{ti}, \overline{A}_{t-1,i})}\right\}. \tag{28}$$

4.1.3 Double-Robust Estimator

According to semi-parametric efficiency theory (e.g., [11, 18]), the efficient score of estimating $\nu(\overline{a}^0)$ is given by

$$D(\nu(\overline{a}^0); P) = \sum_{t=0}^{T} D_t(\nu(\overline{a}^0); P), \tag{29}$$

where P is the true underlying distribution of observation O_i and

$$
\begin{aligned}
D_0(\nu(\overline{a}^0); P) &= Q_0(W_0, a_0^0) - \nu(\overline{a}^0), \\
D_t(\nu(\overline{a}^0; P); P) &= \frac{I[\overline{A}_{t-1} = \overline{a}_{t-1}^0]}{\prod_{s=0}^{t-1} g_s(a_s^0 | \overline{W}_s, \overline{A}_{s-1})}[Q_t(\overline{W}_t, \overline{A}_t) - Q_{t-1}(\overline{W}_{t-1}, \overline{A}_{t-1})], \\
& \qquad t = 1, \ldots, T-1, \\
D_T(\nu(\overline{a}^0; P); P) &= \frac{I[\overline{A}_{T-1} = \overline{a}^0]}{\prod_{t=0}^{T-1} g_t(a_t^0 | \overline{W}_t, \overline{A}_{t-1})}[Y - Q_{T-1}(\overline{W}_{T-1}, \overline{A}_{T-1})].
\end{aligned}
$$

Therefore, if we further obtain estimators of regression functions, $\widehat{Q}_t(h_t, a_t)$ (which can be rewritten as $\widehat{Q}_t(\overline{w}_t, \overline{a}_t)$), then we can obtain the following doubly robust estimator for $\nu(\overline{a}^0)$, by solving the estimating equation $D(\nu(\overline{a}^0); \widehat{P}) = 0$,

$$
\begin{aligned}
\widehat{\nu}_{AIPTW}(\overline{a}^0) = \frac{1}{n}\sum_{i=1}^{n} \Bigg\{ & \frac{I[\overline{A}_i = \overline{a}^0]Y_i}{\overline{g}_{T-1}(\overline{W}_{T-1,i})} + \left[1 - \frac{I[A_{0i} = a_0^0]}{\widehat{g}_0(a_0 | W_{0i})}\right]\widehat{Q}_0(W_{0i}, a_0^0) \\
& + \sum_{t=1}^{T-1}\left[\frac{I[\overline{A}_{t-1,i} = \overline{a}_{t-1}^0]}{\overline{g}_{t-1}(\overline{W}_{t-1,i})} - \frac{I[\overline{A}_{ti} = \overline{a}_t^0]}{\overline{g}_t(\overline{W}_{ti})}\right]\widehat{Q}_t(\overline{W}_{ti}, \overline{a}_t^0)\Bigg\}, \quad (30)
\end{aligned}
$$

where $\overline{g}_t(\overline{W}_{ti}) = \widehat{g}_0(a_0 | W_{0i}) \prod_{s=1}^{t} \widehat{g}_s(a_s^0 | \overline{W}_{si}, \overline{a}_{s-1}^0)$.

4.2 The Standardization Strategy

4.2.1 Estimand

By the standardization strategy, we can define the corresponding estimand for the value of $\overline{a}^0$. That is, under those three identifiability assumptions, $\nu^*(\overline{a}^0)$ is equal to

$$
\nu(\overline{a}^0) = E\{Q_0(W_0, a_0^0)\}, \quad (31)
$$

where regression function Q_0 is defined in (26).

4.2.2 Initial Estimator

If we obtain an estimator of $Q_0(w_0, a_0)$, $\widehat{Q}_0(w_0, a_0)$, then we can obtain the following estimator for $\nu(\overline{a}^0)$:

$$
\widehat{\nu}_{MLE}(\overline{a}^0) = \frac{1}{n}\sum_{i=1}^{n} \widehat{Q}_0(W_{0i}, a_0^0). \quad (32)
$$

4.2.3 Double-Robust Estimator

If we further obtain estimators of propensity-score functions, $\widehat{g}_t$, $t = 0, \cdots, T-1$, we can construct the corresponding doubly robust estimator. For this aim, we apply the backward induction approach. At each decision point $t = T-1, T-2, \ldots, 0$, we first obtain an initial estimator of regression function, $\widehat{Q}_t(\overline{w}_t, \overline{a}_t)$, then we update the initial estimator into $\widehat{Q}_t^*(\overline{w}_t, \overline{a}_t)$ via the targeted learning theory based on the efficient score $D_{t+1}(\nu(\overline{a}^0); P)$, where $\widehat{Q}_t^*(\overline{w}_t, \overline{a}_t)$ is on the least favorable submodel that passes through $\widehat{Q}_t(\overline{w}_t, \overline{a}_t)$. At the end, we obtain $\widehat{Q}_0^*(W_{0i}, a_0^0)$ and thus the doubly robust estimator,

$$\widehat{\nu}_{TMLE}(\overline{a}^0) = \frac{1}{n}\sum_{i=1}^{n} \widehat{Q}_0^*(W_{0i}, a_0^0). \tag{33}$$

4.3 Implementation and Comparison

Similar to Sect. 3.3, here we provide some brief comparison. First, these four methods are generalized from those four methods with the same names in Sect. 3. Second, AIPTW and TMLE are doubly robust versions of IPTW and MLE, respectively. Third, AIPTW relies on parametric modeling of Q_t's and g_t's, while TMLE allows for flexible modeling of Q_t's and g_t's using super learner. Fourth, MLE and TMLE are plug-in estimators, which are more stable than the weighted estimators.

In practice, we can use R package "DTR" [16] to implement $\widehat{\nu}_{IPTW}$, $\widehat{\nu}_{MLE}$, and $\widehat{\nu}_{AIPTW}$, along with their statistical inferences, by specifying GLMs for Q_t's and g_t's. We can use R package "ltmle," with "l" standing for "longitudinal," to implement $\widehat{\nu}_{TMLE}$, by specifying either GLMs or super learner for Q_t's and g_t's. Refer to [22] for a detailed description of R package "ltmle." In the below, we provide an example of using it to estimate the ATE for longitudinal studies with intercurrent events.

Assume that we are interested in estimating the following ATE:

$$\theta = \nu(\overline{1}) - \nu(\overline{0}), \tag{34}$$

which measures the treatment effect of the investigative static treatment regime $\overline{a}^0 = \overline{1}$ compared against the reference static treatment regime $\overline{a}^{0\prime} = \overline{0}$. In order to understand this estimand, we should envisage one hypothetical world in which all the patients follow $\overline{a}^0 = \overline{1}$ throughout the study and the other hypothetical world in which all the patients follow $\overline{a}^{0\prime} = \overline{0}$ throughout the study. That is, in the construction of estimand (34), we apply the hypothetical strategy of ICH E9(R1) [23] to handle intercurrent events (e.g., treatment discontinuation, treatment changing, and rescue medication).

Table 1 The structure of the dataset in one example

Argument	Variable names[a]
Baseline covariates	c("L0.a", "L0.b", "L0.c")
Lnodes[b]	c("L1.a", "L1.b")
Anodes	c("A0", "A1")
Cnodes	c("C0", "C1")
Ynodes	c("Y1", "Y2")

[a] The order of the variables in the dataset: data.frame(L0.a, L0.b, L0.c, A0, C0, L1.a, L1.b, Y1, A1, C1, Y2)

[b] L_t in the Lnodes is the same as W_t in the context

In order to estimate θ in (34), we define the censoring variable C_t, which is a factor variable with two levels, "uncensored" or "censored," at each time t, $t = 0, \cdots, T-1$. If for time t, while $A_s = A_0$ for $s = 0, \cdots, t$, an intercurrent event occurs between t and $t+1$, then $C_t = \cdots = C_{T-1} =$ "$censored$." Note that in this setting we consider the event that directly leads to censoring as the intercurrent event. For example, assume that an adverse event leads to treatment discontinuation, which directly leads to data censoring, and then we consider the treatment discontinuation as an intercurrent event.

To demonstrate the use of R function "ltmle," we look at one example where there are two follow-up visits ($T = 2$), three baseline covariates at $t = 0$ ("L0.a", "L0.b", "L0.c"), two time-dependent covariates at $t = 1$ ("L1.a", "L1.b"), treatment variable measured at $t = 0, 1$ ("A0", "A1"), censoring variable measured at $t = 0, 1$, and outcome variable measured at $t = 1, 2$ ("Y1", "Y2"). Note that the L-node variables form the time-dependent covariates W_t; that is, $W_t = L_t, t = 0, 1$. Table 1 displays the structure of the dataset to be defined in R.

Here is an excerpt of R codes presented in [22] used to implement the TMLE estimator in the above example, providing the point estimate and 95% confidence interval of θ in (34):

```
data <- data.frame(L0.a, L0.b, L0.c, A0, C0, L1.a, L1.b, Y1, A1,
                   C1, Y2)
Lnodes <- c("L1.a", "L1.b")
Anodes <- c("A0", "A1")
Cnodes <- c("C0", "C1")
Ynodes <- c("Y1", "Y2")
ltmle(data = data, Anodes = Anodes, Cnodes = Cnodes, Lnodes = Lnodes,
Ynodes = Ynodes,  survivalOutcome = NULL,
  abar = list(treament = c(1, 1), control = c(0,0)))
```

Here is a remark on how these methods can be extended to survival outcome (a.k.a., time-to-event outcome). In the above R function, "survivalOutcome = NULL" indicates that the outcome variable is either continuous variable or binary having single Ynodes. We set "survivalOutcome = FALSE" for binary outcome variable with multiple Ynodes. For survival outcome, we set "survivalOutcome = TRUE" to indicate that Y_t nodes are indicators of an event, and if Y_t at some time point t is 1, then $Y_s, s = t+1, \cdots, T-1$, should be 1.

5 Discussion

In this chapter, we briefly review some recent statistical development of causal inference methods beyond PS methods. Instead of providing a comprehensive review, we investigate three checkpoints, which may be helpful for guiding us to select an appropriate approach for any study at hand.

If we want to consider one of the four PS methods, then the first checkpoint is whether the conditional approaches (matching, stratification, PS as covariate) or the marginal approach (IPTW). IPTW is a G-method, which can be generalized from point-exposure studies to longitudinal studies.

If we want to consider one of the G-methods, then the second checkpoint is the weighting approaches (e.g., IPTW and AIPTW) or the standardization approaches (e.g., MLE and TMLE). AIPTW is the doubly robust version of IPTW and TMLE is the doubly robust version of MLE.

The third checkpoint is to consider the problem as a time-independent problem or a time-dependent problem. Every G-method has two versions, one simple version for time-independent problem and the other complex version for time-dependent problem. Therefore, all the four methods (IPTW, AIPTW, MLE, and TMLE) have versions for time-dependent problem.

We conclude the chapter with a brief discussion on how to apply these methods to studies with intercurrent events (ICEs). ICH E9(R1) defines ICEs as "events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest. It is necessary to address intercurrent events when describing the clinical question of interest in order to precisely define the treatment effect that is to be estimated." Therefore, we should specify how to handle ICEs in the definition of the estimand and then select an appropriate causal inference method to estimate the estimand.

There are five ICH E9(R1) strategies for handling ICEs: (1) hypothetical strategy, (2) treatment-policy strategy, (3) composite-variable strategy, (4) while-on-treatment strategy, and (5) principal-stratum strategy.

5.1 Hypothetical Strategy for ICEs in Estimand Definition

Applying this strategy, we envision a scenario in which ICEs would not occur and define the estimand of interest as in (34), comparing the treatment regime of taking $A_0 = 1$ throughout against the treatment regime of taking $A_0 = 0$ throughout. The methods described in Sect. 4.3 can be applied to estimate this estimand.

5.2 Treatment-Policy Strategy for ICEs in Estimand Definition

This strategy requires that we collect data even after the ICE occurrence. Applying this strategy, we can use the value of the outcome variable regardless of whether

or not the ICE occurs and define the estimand of interest as in (8) or (14). All the methods described in Sect. 3 can be applied to estimate this estimand without revising the definition of outcome variable.

5.3 *Composite-Variable Strategy for ICEs in Estimand Definition*

Applying this strategy, we need to revise the definition of outcome variable. The new outcome variable is a composite variable of the original outcome variable and the ICE occurrence, and the estimand of interest can be defined as in (8) or (14) with the new outcome variable. All the methods described in Sect. 3 can be applied to estimate this estimand using the new outcome variable.

5.4 *While-on-treatment Strategy for ICEs in Estimand Definition*

Applying this strategy, we need to revise the definition of outcome variable as well. The new outcome variable is a function of the outcome variable measured prior to the ICE occurrence and the time of ICE occurrence (e.g., the rate of change). The estimand of interest can be defined as in (8) or (14) with the new outcome variable. All the methods described in Sect. 3 can be applied to estimate this estimand using the new outcome variable.

5.5 *Principal-Stratum Strategy for ICEs in Estimand Definition*

Applying this strategy, as proposed by ICH E9(R1), "the target population might be taken to be the principal stratum in which an ICE event would occur. Alternatively, the target population might be taken to be the principal stratum in which an ICE would not occur." The estimand of interest can be defined as in (8) or (14), with the outer expectation taken over the principal stratum of interest. To estimate this estimand, we need to estimate the membership of the principal stratum. Then, all the methods described in Sect. 3 can be applied, considering the estimated principal stratum as the target population.

References

1. Greenfield, S., Rich, E.: Welcome to the journal of comparative effectiveness research. *Journal Of Comparative Effectiveness Research*. **1**, 1–3 (2012)

2. Framework for FDA's Real-World Evidence Program, https://www.fda.gov/media/120060/download
3. Fang, Y., Wang, H., He, W.: A statistical roadmap for journey from real-world data to real-world evidence. *Therapeutic Innovation & Regulatory Science*. **54**, 749–757 (2020)
4. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic books, New York (2018)
5. Rubin, D.: Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. **66**, 688–701 (1974)
6. Gelman, A., Vehtari, A.: What are the most important statistical ideas of the past 50 years?. *Journal Of The American Statistical Association*. **116**, 2087–2097 (2021)
7. Rosenbaum, P., Rubin, D.: The central role of the propensity score in observational studies for causal effects. *Biometrika*. **70**, 41–55 (1983)
8. Hernan, M., Robins, J: Causal Inference: What If. Boca Raton: Chapman & Hall/CRC (2020).
9. van der Laan, M.; Rubin, D. Targeted maximum likelihood learning. *The International Journal Of Biostatistics*. **2** (2006)
10. van der Laan, M., Rose, S.: Targeted Learning: Causal Inference for Observational and Experimental Data. Springer (2011)
11. van der Laan, M., Rose, S.: Targeted Learning in Data Science. Springer (2018)
12. Pearl, J.: Causality. Cambridge university press (2009)
13. Austin, P.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*. **46**, 399–424 (2011)
14. Robins, J., Hernan, M., Brumback, B.: Marginal structural models and causal inference in epidemiology. *Epidemiology*. **11** pp. 550–560 (2000)
15. Imbens, G., Rubin, D.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press (2015)
16. Tsiatis, A., Davidian, M., Holloway, S., Laber, E.: Dynamic Treatment Regimes: Statistical Methods for Precision Medicine. Chapman (2020)
17. Fang, Y. Two basic statistical strategies of conducting causal inference in real-world studies. *Contemporary Clinical Trials*. **99** pp. 106193 (2020)
18. Bickel, P., Klaassen, C., Ritov, Y., Wellner, J.: Efficient and Adaptive Estimation for Semiparametric models. Springer (1993)
19. Bang, H., Robins, J.: Doubly robust estimation in missing data and causal inference models. *Biometrics*. **61**, 962–973 (2005)
20. van der Laan, M., Polley, Hubbard, A.: Super learner. *Statistical Applications In Genetics And Molecular Biology*. **6** (2007)
21. Gruber, S., van der Laan, M.: tmle: An R package for targeted maximum likelihood estimation. *Journal Of Statistical Software*. **51** pp. 1–35 (2012)
22. Lendle, S., Schwab, J., Petersen, M., van der Laan, M.: ltmle: an R package implementing targeted minimum loss-based estimation for longitudinal data. *Journal Of Statistical Software*. **81** pp. 1–21 (2017)
23. ICH E9(R1) (2021): Statistical Principles for Clinical Trials; Addendum: Estimand and Sensitivity Analysis in Clinical Trials, https://www.fda.gov/media/148473/download