Targeted Learning: Toward a Future Informed by Real-World Evidence

Susan Gruber a, Rachael V. Phillips b, Hana Lee c, Martin Ho d, John Concato e, and Mark J. van der Laan b

aPutnam Data Sciences, LLC, Cambridge, MA; bDepartment of Biostatistics, School of Public Health, University of California at Berkeley, Berkeley, CA; cOffice of Biostatistics, Center for Drug Evaluation and Research, United States Food and Drug Administration, Silver Spring, MD; dBiostatistics, Google, Mountain View, CA; eOffice of Medical Policy, Center for Drug Evaluation and Research, United States Food and Drug Administration, Silver Spring, MD

ABSTRACT

The 21st Century Cures Act of 2016 includes a provision for the U.S. Food and Drug Administration (FDA) to evaluate the potential use of Real-World Evidence (RWE) to support new indications for use for previously approved drugs, and to satisfy post-approval study requirements. Extracting reliable evidence from Real-World Data (RWD) is often complicated by a lack of treatment randomization, potential intercurrent events, and informative loss to follow-up. Targeted Learning (TL) is a sub-field of statistics that provides a rigorous framework to help address these challenges. The TL Roadmap offers a step-by-step guide to generating valid evidence and assessing its reliability. Following these steps produces an extensive amount of information for assessing whether the study provides reliable scientific evidence, including in support of regulatory decision-making. This article presents two case studies that illustrate the utility of following the roadmap. We used targeted minimum loss-based estimation combined with super learning to estimate causal effects. We also compared these findings with those obtained from an unadjusted analysis, propensity score matching, and inverse probability weighting. Nonparametric sensitivity analyses illuminate how departures from (untestable) causal assumptions affect point estimates and confidence interval bounds that would impact the substantive conclusion drawn from the study. TL’s thorough approach to learning from data provides transparency, allowing trust in RWE to be earned whenever it is warranted.

1. Introduction

The 21st Century Cures Act of 2016 includes a provision for the U.S. Food and Drug Administration (FDA) to evaluate the potential use of Real-World Evidence (RWE) to support new indications for use for previously approved drugs, and to satisfy post-approval study requirements (FDA 2020). The FDA’s draft framework (2018) describes the Real-World Data (RWD) sources that can generate RWE, and outlines challenges and opportunities in developing rigorous approaches to producing reliable evidence (FDA 2018). RWD capture patient health status and health care delivery from a variety of sources, including Electronic Health Records (EHR), medical claims, product and disease registries, patient-generated data, and wearable devices. Analyses of RWD produce RWE representing clinical evidence regarding the usage and potential risks and benefits of medical products. In addition to traditional Randomized Controlled Trials (RCT), other study designs—including pragmatic trials, externally controlled trials, and observational studies—increasingly rely on RWD to produce RWE. Despite legitimate concerns regarding the completeness and accuracy of information on covariates, ascertainment of exposures and outcomes, absence of randomization and loss to follow-up (LTFU), RWD can be a source of valuable insights (Corrigan-Curay, Sacks, and Woodcock 2018; Simon et al. 2022). Considerations for regulatory decision-making include whether the RWD are fit for purpose, whether the study provides adequate scientific evidence, and whether study conduct meets regulatory requirements (FDA 2018, 2021). This article reports on an FDA-funded project to explore the use of Targeted Learning (TL) as a principled approach to incorporating RWD in generating trustworthy evidence.

TL is a statistical framework for efficient learning from data (van der Laan and Rose 2011). The TL estimation roadmap offers step-by-step guidance for transforming a precise statistical question that meets the substantive goal of the study into reliable RWE. The roadmap addresses all attributes of the ICH E9 (R1) Guideline’s definition of a statistical estimand: population, treatment, outcome variable, summary measure, and intercurrent events, defined as “events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest” (ICH 2020; Gruber et al. 2022a). Within the TL framework, Targeted Minimum Loss-based Estimation (TMLE) and Super Learner (SL) are the recommended statistical methodologies for providing efficient estimation and valid inference (van der Laan and Rose 2011). TMLE+SL can appropriately adjust for each source of bias when sufficient confounder information is available.

SL is a vital tool for mitigating model misspecification bias. In high dimensional settings, for example, where rich medical

CONTACT Susan Gruber sgruber@putnamds.com Putnam Data Sciences, LLC, 85 Putnam Avenue, Cambridge, MA 02139.

© 2023 American Statistical Association
histories have been documented in EHR, unbiased effect estimation in finite datasets often requires some form of dimension reduction. With SL, this task can be done sensibly using a variety of data adaptive techniques. Recent applications of SL include risk score prediction (Pirracchio et al. 2015; Gruber et al. 2020), identifying health outcomes of interest (Carrell et al. in press), and estimating causal effects of treatments and exposures in observational and randomized studies (Kreif et al. 2017; Balzer et al. 2019; Kempker et al. 2020).

RCT findings are considered the benchmark research design because causal assumptions (described below) are presumed to be met. However, intercurrent events (such as treatment switching, LTFU, and other forms of right censoring) introduce nonrandomized elements into RCTs that bias unadjusted effect estimates. We demonstrate that TL can be successfully applied to RCTs whether these problems arise or not. Although estimates vary and the situation may be improving with publication of the ICH E9 (R1) Guidelines (ICH 2020), a recent publication found that 99% of RCTs failed to report on missing data separately by outcome, and 95% failed to report a method for judging risk of bias associated with missing data (Kahale et al. 2019). Analysts must remain mindful of this issue even when considering newer studies, particularly those that incorporate historical data, for example, single-arm trials that make use of external controls. When missingness is informative, unbiased estimation requires analyzing the RCT data as if it were observational. It is vitally important that appropriate methods are well understood, and their use becomes mainstream.

This article demonstrates how to integrate TL into evidence generation. We first review the TL Estimation Roadmap, then illustrate its utility in two case studies. In each we evaluated the marginal Additive Treatment Effect (ATE) using data made publicly available by study authors. These analyses and a closely related simulation study illustrate the performance of TMLE+SL compared with the unadjusted estimator, and two popular propensity score (PS) based methodologies—PS matching (Rosenbaum and Rubin 1983) and Inverse Probability Weighting (IPW) (Hernán, Brumback, and Robins 2000). Plasmode simulations based on RCT data from Case study 1 demonstrate how to use TL to overcome common challenges when incorporating RWD. Case study 2 is a pragmatic trial in which treatment was delivered by providers outside the study team. Large loss to follow-up in Case study 2 poses the same challenge to unbiased estimation as is found in studies that incorporate outcomes ascertained from RWD collected for caregiving or billing purposes.

2. The Targeted Learning Estimation Roadmap

The TL estimation roadmap offers a step-by-step guide to estimating causal effects from data (Petersen and van der Laan 2014; Ho et al. 2021). The ICH E9 (R1) attributes of an estimand are considered throughout the roadmap. The population of interest and intercurrent events are part of a prerequisite “Step 0” in which a scientific gap in understanding is expressed in terms of a statistical question that can be addressed by learning from data. In Step 0 domain experts clarify the population; provide precise definitions of treatment and comparator; specify the outcome, length of follow-up, and summary measure of interest; and describe the experiment giving rise to the data. Explicitly considering intercurrent events and adopting amelioration strategies as discussed in the ICH Guidelines is recommended (ICH 2020). For example, consider whether it makes sense to define a composite outcome that includes the outcome event. Another strategy for handling intercurrent events that impact treatment adherence is to consider whether an Intention-to-Treat (ITT) parameter that is independent of adherence addresses the question of interest. For remaining intercurrent events one can develop a plan for collecting baseline and time-dependent covariates that can be used to adjust for potential bias in the estimate of the desired treatment effect.

In our case studies these prerequisites have already been finalized by the study authors. From a regulatory perspective evidence generated from a pragmatic trial using RWD is considered RWE. Thus, we illustrate use of the roadmap from the simplest version of a RWE study (a pragmatic trial) to a highly complex scenario (longitudinal, informative missing, etc.) (FDA 2018; Concato et al. 2020).

Steps 1–3 of the TL roadmap define the statistical and causal models, causal parameter of interest, statistical estimand, and identifying assumptions that link what can be estimated from data with the causal question of interest. This approach naturally integrates intercurrent events and the other ICH-defined elements of a causal estimand. Step 4 of the roadmap involves estimating the statistical parameter and its uncertainty. Step 5 is the conduct of sensitivity analyses to assess the level of support in the data for a causal interpretation of the study finding and substantive conclusion.

Step 1. The statistical model is a collection of possible probability distributions of the data consistent with the data generating experiment that gave rise to the data over time, including intercurrent events. The statistical model should be defined broadly enough to ensure that the true distribution is not excluded. Incorporating domain knowledge, such as known bounds on the outcome, is helpful but the model should not be further restricted based on assumptions not definitively known to be true.

Step 2. The causal parameter, \(\Psi_{\text{causal}}\), is a feature of the true data distribution, defined in terms of a causal model of the full (counterfactual) data, rather than as a coefficient in a particular parametric model. The causal model expresses conditional independencies, assumes exogenous errors, and should be in alignment with the statistical model. Directed Acyclic Graphs (DAG) are often a useful tool for depicting these relationships. Examining their structure can help identify sufficient sets of confounders to adjust for in the analysis (Pearl 2009; Hernán and Robins 2020).

Step 3. The statistical estimand, \(\Psi_{\text{obs}}\), is a quantity that can be estimated from observed data that most closely approximates \(\Psi_{\text{causal}}\). \(\Psi_{\text{obs}}\) and \(\Psi_{\text{causal}}\) are linked through identifying assumptions informally known as (i) positivity, an assumption that within strata defined by confounders there is a nonzero probability of receiving treatment at all levels of interest; (ii) consistency, an assumption that the outcome observed under the assigned treatment is equivalent to the counterfactual outcome under that level of exposure;
and (iii) Coarsening at Random (CAR), which implies no unmeasured confounding. PS diagnostics (e.g., summaries of covariate balance, distribution of the PS in each treatment arm and/or IP weights, C-statistic) can indicate whether the positivity assumption is met, but the other two assumptions are not testable from data. When all three assumptions hold, $\Psi^{obs}$ is an expression of the causal quantity in terms of observable rather than counterfactual data. Alternative sets of identifying assumptions may be sufficient in some settings (Pearl 2010; Hernán and Robins 2020).

To a certain extent Steps 1–3 are part of an iterative process. After precisely specifying what we want to learn and characterizing the data generating process in step 0, we grapple with the quality of the available data in light of the causal and statistical model specifications, and the plausibility of identifying assumptions. If the causal and statistical models are not in agreement or insufficient confounder information is available one might modify the target parameter (in terms of the summary measure, target population, treatment or outcome specification), or consider the plausibility of additional/alternate assumptions that would improve identifiability. Dissatisfaction might lead study authors to revisit Step 0 to formulate more attainable study goals.

**Step 4.** TMLE+SL is recommended for estimation of $\Psi^{obs}$, due to its robust finite sample performance and flexibility in addressing the challenges inherent in analyses of RWD. SL provides data adaptive modeling that mitigates model misspecification bias. TMLE respects bounds on the problem, and unlike PS-based methodologies, uses all available information in the data. This typically results in estimates with smaller Mean Squared Error (MSE) than PS-based methods (Porter et al. 2011; Lendle, Fireman, and van der Laan 2013). Because the sampling distribution of the TMLE is well understood, one can obtain valid inference, that is, estimates of uncertainty, $p$-values, and 95% Confidence Intervals (CI) with good coverage properties.

**Step 5.** Sensitivity analyses help assess the robustness of conclusions drawn from the study findings. A suite of sensitivity analyses can build a rich picture of the reliability of a study finding. TL prescribes a nonparametric investigation of the impact on results of any causal gap, the nonrandom portion of the difference between the true quantity of interest and its estimated value due to departures from the identifying assumptions (Diaz and van der Laan 2013). Domain knowledge and an understanding of the experiment that gave rise to the data can provide insight into the magnitude of realistic departures from the underlying assumptions. Understanding how point estimates and CIs are affected indicates the level of support in the data for the substantive conclusion and the reliability of the RWE.

Analogous to the recently proposed E-value which quantifies the minimum strength of association (on the risk ratio scale) that an unmeasured confounder would need to have with both the exposure and the outcome to fully explain away a specific observed exposure-outcome association conditional on measured covariates (VanderWeele and Ding 2017), we define the G-value as the minimal causal gap that would negate the study finding (i.e., cause the 95% CI to include the null if it is currently excluded, or exclude the null if it is currently included). Under this definition the G-value is equal to the minimum distance from the null value to either the upper or lower bound of the 95% CI. The G-value can be equivalently expressed in “causal gap” units on the same scale as the effect estimate, in “SE units” relative to the Standard Error (SE) of the estimate, or in “Adj units” relative to the difference between the adjusted and unadjusted estimates. For both the E-value and the G-value subject matter experts can consider whether the value falls into a plausible range. Unlike the G-value, the E-value focuses on contributions to the causal gap from potential violations of only one of the identifying assumptions, ignoring contributions from the others, thus, we recommend it not be considered in isolation.

### 3. Estimators of the Additive Treatment effect (ATE)

Consider data consisting of iid observations, $O = (\Delta Y, \Delta, A, W)$, where $\Delta Y$ equals the outcome value when $\Delta = 1$ and is missing when $\Delta = 0$, $\Delta$ is a binary indicator of outcome missingness, $A$ is a binary treatment indicator, and $W$ is a vector of baseline covariates. The likelihood of the data can be factorized as $L(O) = P(Y|\Delta, A, W), P(\Delta|A, W), P(A|W), P(W)$. The TMLE literature refers to $P(Y|\Delta, A, W)$ and $P(W)$ as the Q portion of the likelihood, and to $P(\Delta|A, W), P(A|W)$ as the G portion of the likelihood (van der Laan and Rose 2011). The ATE parameter is defined in terms of the Q portion of the likelihood, $\psi_{ATE}^{causal} = E(Y_1) - E(Y_0)$, where $Y_a$ is the counterfactual conditional mean outcome under exposure at level $A = a$, and the expectation is with respect to the distribution of $W$. When identifying assumptions are met, the corresponding statistical estimand in the observed data is given by $\psi_{ATE}^{obs} = E(Y|\Delta = 1, A = 1, W) - E(Y|\Delta = 1, A = 0, W)$.

We estimated $\psi_{ATE}^{obs}$ using an unadjusted estimator, several variants of full PS matching, Inverse Probability of Treatment Weighting (IPTW) (or Inverse Probability of Treatment and Censoring Weighting (IPTCW), when the data contained missing outcomes), and TMLE+SL. Each estimator’s efficiency depends on the information in the data. The unadjusted estimator of the treatment effect uses a subset of information available in the Q portion of the likelihood to estimate $\psi_{ATE}^{obs}$. PS matching uses information in only the PS component of $G$, $P(A|W)$. IPTW estimators use information in both components of $G$. TMLE incorporates information from all components of $Q$ and $G$. Using more information in the data decreases asymptotic variance. Implementation details for each estimator are described next. All analyses were run in the R statistical programming environment, v3.6.1 (R Core Team 2020). Sample code is available as supplementary materials.

#### 3.1. Propensity Score Matching (PS Matching)

Although PS matching is often used to estimate an average treatment effect among the treated, full matching can also be used to evaluate the ATE (Hansen 2004). Full matching on the logit of the PS was carried out using the MatchIt package, v3.0.2 (Ho et al. 2011). MatchIt produces weights proportional to each observation’s contribution to one or more matched sets. These weights were incorporated into an unadjusted logistic regression
of the outcome on treatment (variant 1), and as recommended in practice (Stuart 2010), an adjusted regression of the outcome on treatment and all covariates in the dataset (variant 2). Post-match adjustment is also recommended practice (Stuart 2010). The ATE estimates were evaluated by calculating the mean difference in predicted probabilities of experiencing the outcome when treated versus not treated.

For each matching variant, the PS was estimated using a main terms logistic regression model. As matches only consider the PS, ignoring the outcome, observations with missing outcomes were retained as potential matches, but then dropped from the subsequent analysis. Although common practice, and we use such PS matching estimators for comparison purposes, this approach should not be construed as the ideal version of matching. When poor overlap in treated and comparator populations results in many dropped observations, it can be difficult to understand among whom the treatment effect has been estimated, and to whom it might apply. More recently proposed matching estimators, for example, those that incorporate multiple imputation or that use machine learning to fine-tune the propensity score estimates to reduce covariate imbalances (Ramsahai, Grieve, and Sekhon 2011; Imai and Ratkovic 2013; Ling et al. 2019), were not considered.

### 3.2. Inverse Probability Weighting (IPW)

IPW estimates of the ATE were obtained from a weighted regression of the outcome, \( Y \), on the binary treatment indicator, \( A \). This univariate regression will yield the identical value for the ATE, though not the model coefficients, regardless of whether logistic or linear regression is used. We used linear regression, because the ingredients for evaluating a robust SE for the ATE estimate (the coefficient in front of treatment in the model) were readily available. Robust SEs were calculated using the \texttt{sandwich} package v2.5-1 (Zeileis 2004; Zeileis, Koll, and Graham 2020).

Stabilized IP weights have been recommended over their unstabilized counterparts for reducing finite sample bias and variance (Hernán, Brumback, and Robins 2000). These weights were calculated as

\[
wt_{\text{stab}} = \frac{\Delta}{P(A = 1|A, W)} \times \left[ \text{mean}(A) \frac{A}{\hat{P}_S} + \text{mean}(1 - A) \frac{1 - A}{1 - \hat{P}_S} \right].
\]

When no observations are missing, this general formula reduces to the familiar IPTW equation for stabilized weights inside the square brackets. The IPW estimator is undefined when any term in the denominator of \((1)\) equals zero, that is, the positivity assumption is violated. Otherwise, an observation could receive an infinite IP weight. Recommended practice is to place on upper bound on the maximum allowed weight, to avoid overly large contributions from only a few influential observations (Cole and Hernán 2008). We investigated four truncation strategies, truncating the weights at either 40, 100, 1,000,000 (essentially, no truncation), and our recently proposed sample size-based bound of \(\sqrt{n} \ln(n)/5\) (Gruber et al. 2022b). For each strategy, any weight that exceeds the specified value is set to the allowed maximum.

The PS and missingness probabilities were estimated using a main terms logistic regression model regressing treatment on all available covariates (variant 1) or using SL (variant 2).

### 3.3. Targeted Minimum Loss-based Estimation with Super Learner (TMLE + SL)

TMLE analyses were carried out using the \texttt{tmle} R package, v1.5.0-1 (Gruber and van der Laan 2012), and the \texttt{SuperLearner} R package, v2.0-24 (Polley et al. 2019). Except where stated, we used each package’s default specifications. In particular, the SL library for modeling the outcome regression included linear regression, Bayesian additive regression trees (BART, in \texttt{dbarts} v0.9-11) (Chipman, George, and McCulloch 2010), and lasso (\texttt{glmmnet} v3.0-2) (Friedman, Hastie, and Tibshirani 2010). The library for modeling the PS and missingness mechanisms (G components) included logistic regression, BART, and generalized additive models (\texttt{gam} v1.16) (Hastie 2019).

Analogous to IPW, TMLE requires the product of the G components of the likelihood to be bounded away from zero. We evaluated performance under four different PS truncation levels, equal to the reciprocal of the bounds on the IP weights: 0.025, 0.01, 0.000001, \(5\left[\sqrt{n} \ln(n)\right]\).

### 4. Case Study 1: A Traditional Randomized Controlled Trial with Minimal Loss to Follow-Up

Case study 1 illustrates how to apply each step of the roadmap and establishes baseline performance characteristics of each estimator in the context of a nearly ideal RCT (e.g., large number of randomized patients with minimal or noninformative LTFU and treatment nonadherence). An additional plasmode simulation study (Franklin et al. 2014) provides a comparison of performance under simulated selection bias and informative LTFU.

The International Stroke Trial (IST) was a multi-national RCT conducted in 1991–1996 to test the safety of aspirin and heparin in patients hospitalized for stroke (International Stroke Trial Collaborative Group 1997). Patients suspected of suffering an ischemic stroke (\(n = 19,435\)) were randomized to aspirin or no aspirin, with and without heparin, to assess several outcomes of interest. In an ITT analysis the authors reported a statistically significant effect of treatment with aspirin versus no aspirin on recurrent ischemic stroke within 14 days after randomization (\(\psi^{ATE}_0 = -0.011, p < 0.001\)). The authors reported a non-statistically significant effect of treatment with aspirin versus no aspirin on mortality within 14 days after randomization (\(\psi^{ATE}_0 = -0.004, p > 0.05\)). De-identified data were downloaded from an online repository (Sandercock, Niewada, and Czlonkowska 2011).

### 4.1. Complete Case Data analysis

The complete case analytic dataset contained \(n = 19,408\) observations. Dropping 27 observations where either outcome was missing mimics what was done in the original study and does not meaningfully alter the data distribution, the definition of the target population, or the magnitude of the effect estimate.
The data structure when all outcomes are observed simplifies to \( O = (Y, A, W) \), where \( W \) consists of 24 continuous, binary, and categorical covariates (Supplemental Appendix Table A1).

### 4.1.1. Applying the Targeted Learning Roadmap

**Step 1.** Statistical model: All distributions of the data consistent with study inclusion/exclusion criteria, respecting the process that gave rise to the data. The likelihood of the data can be factorized as \( P(Y|A, W)P(A|W)P(W) \). We place no restrictions on the functional form of the relationships between treatment, covariates, and the outcome other than to note that \( Y \) is bounded between 0 and 1. Contrast this with a main terms logistic regression model that imposes a monotonicity assumption on covariate-outcome relationships, and assumes no effect modification or treatment effect heterogeneity. Because treatment was randomized, we have considerable knowledge about the assignment mechanism. There is no systematic driver of treatment, though there may be chance imbalances. Theory recommends efficiency can be improved by modeling treatment as a main terms logistic regression that includes covariates associated with the outcome (Rosenblum and van der Laan 2010). Using a parametric regression model guarantees an achieving an optimal rate of convergence. The empirical distribution of \( W \) offers a consistent estimate of the final factor.

**Step 2.** Causal model and causal parameter: The causal parameter of interest reported in the original published study is the ATE, \( \Psi_{\text{causal}}^{ATE} = E(Y_1) - E(Y_0) \). Our causal model assumes exogenous errors, and independencies implied by the time ordering.

**Step 3.** Statistical estimand and identifiability: The statistical estimand is given by \( \Psi_{\text{obs}}^{ATE} = E(Y|A = 1, W) - E(Y|A = 0, W) \). Knowledge of the process that gave rise to the data aids in assessing the plausibility of the identifying causal assumptions. (i) The positivity assumption is guaranteed by baseline randomization, and by the nearly complete follow-up. (ii) Adherence to assigned treatment was 88% or more in each treatment arm, with cessation largely due to hospital discharge. Under a slightly modified definition of treatment, treat within 48 hr of stroke, and continue until hospital discharge, 14-day adherence is essentially 100%. Thus, the consistency assumption is satisfied. (iii) Baseline randomization, adherence to treatment, and complete follow-up imply the CAR assumption is satisfied.

**Step 4.** Estimation: An unadjusted logistic regression, PS matching, IPTW, and TMLE+SL were used to estimate the ATE. For PS matching, we enforced an exact match on region by stratifying the matching procedure. For IPTW, the PS was estimated using a main terms logistic regression model, in which treatment was regressed on all available baseline covariates. For TMLE, SL was used for modeling the outcome, but we specified a parametric logistic regression model for the PS instead of using SL, because when analyzing RCT data theory guarantees consistency of the TMLE when PS estimates are modeled parametrically. The covariates incorporated as main terms in the PS model for the TMLE were prescreened using a lasso algorithm that retains covariates associated with the outcome. Note that the automated screening process is an internal part of the TMLE estimation procedure, and thus adheres to the outcome blinding principle (for the analyst).

**Step 5.** Sensitivity analyses and substantive conclusion: The sensitivity analysis is designed to illustrate the point estimate and CI bounds for the causal parameter under hypothesized asymptotic distances between the causal and statistical estimands. This causal gap reflects the combined impact of violations of the identifying assumptions, including misspecification of the statistical model. The first step is to quantify how the value of estimates and CI bounds change under causal gaps of different magnitudes and direction. To facilitate comparison the magnitude can be expressed in the same units as the original estimate (\( \Lambda \)), in terms of the SE of the original estimate (SE units), relative to the difference between the unadjusted and adjusted estimate (Adj units), or another meaningful metric. In addition to calculating the G-value, looking at the causal gap across a range of values provides a comprehensive picture illustrating the gap size required for the CI to include the null, for it to completely exclude the null, and for it to completely exclude the null in the other direction (reversing the conclusion one would draw from the study). This information is most easily conveyed in a plot.

The next step is to critically consider the potential sources of bias and come to an expert or external data-driven assessment of the direction of their combined impact, and plausible bounds on the magnitude. For example, in an ideal RCT all three identifying assumptions are trivially met. In this case study randomization and negligible LTFU ensures the positivity assumption and CAR assumptions are met. The consistency assumption is guaranteed in any ITT analysis, assuming accurate outcome assessment and record keeping. Thus, the expected causal gap is presumably quite close to zero. Results of the sensitivity analysis are presented in Section 4.1.3.

### 4.1.2. Results

**Propensity score diagnostics.** Common diagnostics allow us to understand how closely the treatment and control groups resemble each other. These diagnostics include a plot of the PS distribution within each treatment group showing near perfect overlap (Figure 1), the distribution of IP weights, and the standardized mean difference for all covariates before and after matching and IP weighting (Supplemental Appendix Table A2). Because randomization was successful, there were no imbalances before or after matching and IP weighting. The C-statistic of 0.51 indicates that the covariates are not predictive of treatment assignment. Stabilized IP weights were between 0.78 and 1.35 for all observations, with a mean of 1.0. Truncation at the levels we considered had no effect on the weights.

**Estimator performance.** Estimates, SEs, 95% CIs, and p-values based on analytic estimates of the variance, as well as bootstrapped variance estimates for unadjusted, TMLE, and IPW estimates, are reported for the 14-day mortality outcome and the 14-day recurrent stroke outcome (Table 1). Because no IP weights or PS values were modified by truncation at any of the levels considered, only one IPW result and one TMLE result are displayed in the table.
Figure 1. Subtask 1.1, Propensity score distribution in treated and control groups in the International Stroke Trial.

Table 1. Marginal additive treatment effect (ATE) point estimates, standard errors (SE), p-values, 95% confidence interval lower bounds (lb) and upper bounds (ub) based on analytic variance estimates and bootstrapped variance estimates for analyses of 14-day mortality and 14-day recurrent ischemic stroke outcomes.

| Estimator                  | Estimate | SE  | p-value | lb     | ub     | SE  | p-value | lb     | ub     |
|----------------------------|----------|-----|---------|--------|--------|-----|---------|--------|--------|
| 14-day mortality\(^a\)     | −0.004   | 0.004| 0.381   | −0.011 | 0.004  | 0.004| 0.391   | −0.012 | 0.005  |
| Unadjusted                 | −0.004   | −    | −       | −      | −      | 0.005| 0.369   | −0.013 | 0.005  |
| Matching unadj             | −0.003   | −    | −       | −      | −      | 0.004| 0.518   | −0.012 | 0.006  |
| Matching adj               | −0.003   | 0.004| 0.407   | −0.011 | 0.005  | 0.004| 0.388   | −0.011 | 0.004  |
| IPTW                       | −0.003   | 0.004| 0.575   | −0.011 | 0.006  | 0.004| 0.530   | −0.010 | 0.005  |
| 14-day recurrent ischemic stroke\(^b\) | −0.009   | 0.002| < 0.001 | −0.013 | −0.005 | 0.002| < 0.001 | −0.013 | −0.005 |
| Unadjusted                 | −0.009   | −    | −       | −      | −      | 0.002| < 0.001 | −0.014 | −0.004 |
| Matching unadj             | −0.009   | −    | −       | −      | −      | 0.002| < 0.001 | −0.013 | −0.004 |
| Matching adj               | −0.009   | −    | −       | −      | −      | 0.002| < 0.001 | −0.013 | −0.005 |
| IPTW                       | −0.009   | 0.002| < 0.001 | −0.013 | −0.005 | 0.002| < 0.001 | −0.013 | −0.005 |
| TMLE                       | −0.009   | 0.002| < 0.001 | −0.013 | −0.005 | 0.002| < 0.001 | −0.013 | −0.005 |

NOTE: Estimators include an unadjusted logistic regression model (Unadjusted); propensity score matching (Matching) without post-match adjustment (unadj), or with post-match adjustment (adj); inverse probability of treatment weighted (IPTW); and targeted minimum loss-based estimation (TMLE).

\(^a\)Reported ATE for mortality = −0.004.

\(^b\)Reported ATE for stroke outcome = −0.011.

All causal methodologies produced results quite similar to the unadjusted estimate, and to published findings. Treatment had essentially no impact on 14-day mortality but conferred slight protection against 14-day recurrent stroke. Minor differences from published results may stem from the fact that our dataset contained slightly fewer observations due to missing outcomes. Bootstrap estimates of the SEs closely agree with analytic SE estimates (not available for the post-match adjusted matching estimator), and all estimators have similar SEs. Because the outcome is rare there is little variation in the outcome to explain, so little opportunity for variance reduction.

4.1.3. Sensitivity Analysis and Substantive Conclusion

Since all three identifying assumptions are met by the study’s design and conduct, the only possible source of any causal gap is that our complete case analysis omitted 27 of the 19,435 observations in the data. The most extreme bias would be observed if either all omitted observations in the treatment arm were true cases and all omitted observations in the control arm were non-cases, or the reverse. We therefore created these two hypothetical versions of the data and carried out an unadjusted regression of the outcome on treatment. The point estimates, SE, and 95% CIs indicate there is essentially no change in the point estimate and 95% CI at the most extreme plausible values for the causal gap (Table 2).

The G-value equals the upper bound on the 95% CI minus the null value of 0: 0.006 causal gap units, 0.006/0.004 = 1.5 SE units, 0.006/0.001 = 6 Adj units. Figure 2 illustrates that with respect to 14-day mortality (left), the lower bound on the 95% CI is negative under most of the plausible values of the causal gap investigated (“Adj Units” are the ratio of the causal gap to the amount of adjustment due to measured covariates). This indicates strong support for a conclusion that treatment does not increase risk of mortality. Because the CI around the actual study finding includes the null and there is little reason to think the causal gap is large, the sensitivity analysis also supports the conclusion that treatment has no impact on mortality. However, it provides essentially no support for a conclusion that treatment increases risk of mortality; for that inference to be true, the
causal gap would have to be 3.4 times as large as the adjustment due to measured confounders.

With respect to 14-day recurrent stroke the G-value is equal to the absolute value of the upper bound on the 95% CI: 0.005 causal gap units, 2.5 SE units, undefined in terms of Adj units of size 0. The full sensitivity analysis (Figure 2, right) clearly supports the conclusion that treatment decreases risk for stroke. The causal gap would have to be nearly 70% larger than the effect size for the 95% CI to include the null. Recall that adjusting for potential baseline confounders had no impact on point estimates, treatment was randomized, and there was nearly no LTFU. In this well designed and well conducted study there are no realistic contributors to a causal gap of that magnitude. The sensitivity analyses demonstrate strong support for the substantive conclusions originally reported.

4.2. Plasmode Simulation Studies

Three plasmode simulation studies were designed to investigate estimator performance when informative selection into treatment and informative LTFU introduce bias into the unadjusted causal effect estimate. The data generating procedure is described in detail in the Appendix. Briefly, for each simulation study we generated 1000 datasets by drawing \( n = 19,435 \) observations with replacement from the original IST data, then assigning approximately 15% to the treatment group (\( A = 1 \)) conditional on covariates, and finally generating a binary outcome \((Y)\) conditional on treatment and covariates. Prior to analyzing the data we imposed informative missingness on the outcomes. In plasmode study I binary missingness indicators, \( \Delta \), were generated conditional on treatment and baseline covariates such that the outcome was observed \((\Delta = 1)\) in approximately 95% of observations, and erased from the data (set to "NA" in R) in the remaining 5% of observations having \( \Delta = 0 \). The proportion of missing outcome values was increased to 15% and 25% in plasmode studies II and III, respectively. In all studies the true ATE is \( \psi^{ATE} = -0.1381 \).

4.2.1. Applying the Targeted Learning Roadmap

Definitions of the statistical model, causal model, target causal parameter, and statistical estimand, defined in Steps 1–3 of the roadmap, remain unchanged. However, the Step 3 assessment of the identifying assumptions linking the observed data to the full data need to be reexamined in light of selection bias and LTFU. For observations where the outcome is not missing consistency is met. PS diagnostics help determine the plausibility of the positivity assumption (results below). As the designers of the simulation we can guarantee that the CAR assumption is met, but an independent data analyst would have no such assurance.

Estimation in Step 4 was carried out using generalizations of each estimator considered in the case study. Matching is an outcome-blind procedure, so all observations were eligible to be matched. After the matches were established, all observations where the outcome was missing were dropped from the dataset.
Both IPW variants described in Section 3.2 were extended that is, estimating the PS and missingness mechanism with logistic regression versus SL. For TMLE, SL was used to estimate the PS, missingness mechanism, and outcome regression, using the default SL specifications in the tmle software.

4.2.2. Results

Propensity score diagnostics. These are derived from a single dataset consisting of covariates from the original observations, and simulated treatment. The plot of PS distributions within each treatment group shows that although there are some differences between the two groups, there is considerable overlap (Figure 3). The C-statistic of 0.78 supports this conclusion.

Standardized mean differences in the observed data indicate that covariate imbalances are greatly reduced after matching or IP weighting (Supplemental Appendix Table A4). The IPTC weights account for imbalances due to both selection into treatment and missingness. The stabilized IPTC weights ranged between 0 and 16 in all studies, so truncation again had no impact on the results (Table 3). As expected, the mean stabilized weight is approximately equal to the proportion of observations where the outcome is observed. Although it seems counterintuitive that the maximum observed weight is largest when only 5% of observations are missing, and smallest when 25% are missing, this situation occurred because observations where subjects happened to receive rare treatment assignments tended to be missing more often. When the outcome is missing, the observation is assigned a weight of zero. These diagnostics suggest the positivity assumption is met. In this simulation study the validity of the randomization assumption is known by design, but in practice it cannot be guaranteed when treatment is not randomized or when there is LTFU.

Estimator performance. Mean bias, standard deviation of the bootstrap estimates (SE), MSE, as well as coverage of 95% CI based on the bootstrap SE and the estimated analytic SE for all three simulation studies, are shown in Table 4. PS Matching was less biased than the unadjusted estimate, but model misspecification and inappropriate handling of missing outcomes produced biased estimates with large SEs. Post-match adjustment greatly reduced the MSE, and somewhat improved the CI coverage. IPTCW that relied on parametric modeling was fairly successful, but incorporating SL greatly reduced bias and slightly reduced the SE. This more favorable ratio of bias to SE manifests as greatly improved CI coverage (Gruber et al. 2022b). In all scenarios, TMLE had the smallest bias and MSE, and nearly optimal CI coverage.

5. Case Study 2: Pragmatic Trial with 25% Loss to Follow Up

The Acupuncture for Chronic Headache in Primary Care trial (ACHPC) was a pragmatic trial carried out in England and Wales in 1999–2001 to assess the effect of acupuncture in practice on headache (Vickers et al. 2004). The goal was to determine the effects of a policy of “use acupuncture” and use of resources in patients with chronic headache compared with a policy of...
“avoid acupuncture.” De-identified data were made publicly available by the study team (Vickers 2006). Adult general care patients having chronic headache (n = 401) were randomized to either 3 months of acupuncture (12 treatments), or usual care. Randomization was site-specific, conditional on age, sex, headache type, baseline headache score, and number of years of headache disorder. Randomization probabilities were calculated for each subject to correct for chance imbalances among subjects who were previously randomized (Vickers et al. 2004). The target statistical estimand was the ITT effect of acupuncture versus no acupuncture on headache score, an ATE. Unbiased causal effect estimation was complicated by LTFU. The primary outcome measure, mean weakly headache score (MWHS) at the 12-month end of study, was missing in 25% of observations. Study authors reported results from a complete case analysis that adjusted for baseline determinants of randomization probabilities, and from sensitivity analyses where multiple imputation was used to generate missing outcome values. In a separate publication, the authors demonstrated a clear effect of treatment randomization on a correlated outcome, a global headache metric, that was available for 95% of subjects (Vickers and McCarney 2003). We focus on the original MWHS outcome and consider a longitudinal data analysis, to illustrate how the TL roadmap informs causal inference when outcomes are subject to right censoring, and to compare estimator performance in this setting.

5.1. Estimating the ATE: Longitudinal Data analysis

The dataset consisted of n = 401 iid observations, and outcomes were missing for 100 (25%) of observations. These publicly available data provide an opportunity to apply methodologies that were not available at the time of the original analysis (TMLE+SL and IPW), and to evaluate the relative performance of TMLE, IPW, and PS Matching when there is substantial LTFU. Study authors collected data on medications, headache score, and care usage periodically throughout the study. We can exploit the availability of these time-varying covariates by analyzing the data from a longitudinal perspective. Data were collected at 3, 5, 9, and 12 months. Right censoring also occurred at these time points. Let L denote censoring nodes along the timeline (Figure 4). The set {A0, A1, A2, A3, A4} refers to the set of intervention nodes, because the causal contrast of interest involves intervening to assign a specific counterfactual treatment and to ensure there is no right censoring.

Patients LTFU were slightly younger, had higher baseline headache scores, and were more likely to be in the control arm (22% treated vs. 29% control). Three covariates had missing values (pain at baseline, physical functioning, and role limitation physical at baseline, role limitation emotional at baseline). Prior to analyzing the data missing covariate values were imputed by

Table 4. Mean bias, standard deviation of the bootstrap estimates (SE), mean squared error (MSE), and coverage of 95% confidence intervals (CI) based on the bootstrap SE and the estimated analytic SE for simulation studies I, II, and III.

| Estimator       | Bias  | SE    | MSE   | 95% CI Coverage | Bootstrap SE | Analytic SE |
|-----------------|-------|-------|-------|-----------------|--------------|-------------|
| I. 5% Missing   |       |       |       |                 |              |             |
| Unadjusted      | 0.0722| 0.0065| 0.00526| 0               | 0            |             |
| IPTCW           |       |       |       |                 |              |             |
| SL              | 0.0201| 0.0050| 0.00003| 0.948           | 0.944        |
| II. 15% Missing |       |       |       |                 |              |             |
| Unadjusted      | 0.0736| 0.0067| 0.00546| 0               | 0            |             |
| IPTCW           |       |       |       |                 |              |             |
| SL              | 0.0201| 0.0053| 0.00003| 0.934           | 0.962        |
| III. 25% Missing|       |       |       |                 |              |             |
| Unadjusted      | 0.0755| 0.0071| 0.00574| 0               | 0            |             |
| IPTCW           |       |       |       |                 |              |             |
| SL              | 0.0202| 0.0058| 0.00004| 0.934           | 0.960        |

NOTE: Estimators include an unadjusted logistic regression model (Unadjusted); propensity score matching (Matching) without post-match adjustment (unadj), or with post-match adjustment(adj); inverse probability of treatment and censoring weighted (IPTCW), with propensity score and missingness mechanisms estimated using logistic regression (GLM) or super learner (SL); targeted minimum loss-based estimation (TMLE).
assigning the mode for binary variables, and the median for continuous variables. Corresponding binary missingness indicators were created and added to the dataset.

### 5.1.1. Applying the TL Roadmap

**Step 1.** Statistical model: All distributions of the data consistent with study inclusion/exclusion criteria, respecting the process that gave rise to the data. The data likelihood factorizes as \( \prod_{t=1}^{T} P(L_t|L_{t-1}, \bar{A}_t) P(A_t|\bar{L}_0) P(\bar{L}_0) \). No restrictions are placed on the functional form of the relationships, however, the distribution of the outcome included in the statistical model must respect known lower and upper bounds on the MWHS of 0 and 70. Treatment and censoring probabilities are bounded between 0 and 1. In addition, treatment was known to be randomly assigned at baseline conditional on six covariates, so a parametric model that conditions on those known potential confounders and other predictors of the outcome is a sound option in finite samples that guarantees asymptotic linearity (van der Laan and Robins 2003).

**Step 2.** Causal model and causal parameter: The causal parameter of interest is the effect of randomization to acupuncture (A_0 = 1) versus usual care (A_0 = 0), under no right censoring. The target parameter is defined as, \( \Psi_{ATE} = E(Y_{11111}) - E(Y_{01111}) \), in a causal model where baseline covariates, treatment, intercurrent events, time-varying covariates, and the outcome are explicitly represented. The subscripts denote the counterfactual settings for each intervention node (baseline treatment, \( A_0 \), and four missingness indicators). This definition of the parameter explicitly acknowledges the potential for intercurrent events at some time \( t > 0 \), to introduce bias. The causal model assumes exogenous errors, and only known conditional independencies or those implied by the time ordering.

**Step 3.** Statistical estimand and identifiability: The statistical estimand is given by \( \Psi_{ATE} = \Psi_{ATE1} - \Psi_{ATE0} \), where \( \bar{A}_1 \) is the longitudinal intervention that sets the intervention nodes to (1,1,1,1,1), and \( \bar{A}_0 \) is the longitudinal intervention that sets the intervention nodes to (0,1,1,1,1). By convention, \( \bar{A}_t \) denotes the intervention node settings through time \( t \), and \( \bar{L}_t \) denotes the covariate history through time \( t \). \( \bar{Q}^{\Psi_1} \) and \( \bar{Q}^{\Psi_0} \) can each be expressed as a series of nested conditional means, \( \bar{Q}^{\Psi} = \left( \bar{Q}^{\Psi}_{Y}, \bar{Q}^{\Psi}_{L_4}, \bar{Q}^{\Psi}_{L_3}, \bar{Q}^{\Psi}_{L_2}, \bar{Q}^{\Psi}_{L_1}, \bar{Q}^{\Psi}_{L_0} \right) \) (Bang and Robins 2005; Petersen et al. 2014),

\[
\begin{align*}
\bar{Q}^{\Psi}_{Y} &= E(Y|L_4, \bar{A}_4 = \bar{a}_4), \\
\bar{Q}^{\Psi}_{L_4} &= E(L_4|L_3, \bar{A}_3 = \bar{a}_3), \\
\bar{Q}^{\Psi}_{L_3} &= E(L_3|L_2, \bar{A}_2 = \bar{a}_2), \\
\bar{Q}^{\Psi}_{L_2} &= E(L_2|L_1, \bar{A}_1 = \bar{a}_1), \\
\bar{Q}^{\Psi}_{L_1} &= E(L_1|L_0, A_0 = a_0), \\
\bar{Q}^{\Psi} &= E(\bar{Q}^{\Psi}_{L_1}).
\end{align*}
\]

Assessment of identifiability: The sequential positivity assumption must hold at all timepoints. Baseline treatment randomization guarantees positivity holds at \( t = 0 \), but not necessarily at subsequent timepoints. If right censoring differentially depletes a large proportion of observations within strata of confounders by trial arm, then a practical violation of the positivity assumption can occur. Time-varying PS diagnostics can provide helpful information for assessing whether this assumption is met. For the ITT analysis, the consistency assumption is easily assessed as satisfied, since adherence to assigned treatment is immaterial. The CAR assumption is satisfied with respect to treatment assignment but is potentially violated if there are unmeasured confounders of the associations between treatment, right censoring, and the outcome.

**Step 4.** Estimation: The statistical parameter can be estimated with longitudinal versions of IPTW (L-IPTW) and TMLE (L-TMLE), available in the ltmle R package (v. 1.2-0) (Lendle et al. 2017). Both estimators require modeling the PS and conditional probabilities of being observed at each intervention node, given the past (\( G_t \)). L-TMLE evaluates a nested conditional mean outcome regression at each time point, targeted toward the parameter of interest based on the cumulative product of the components of \( G \) from \( t_0 \) through \( t \). The longitudinal IP weight at time point \( t \) is the cumulative product of the components of \( G \) from \( t_0 \) through \( t \). Weights were truncated using the sample size-based strategy, \( w_{\text{max}} = \sqrt{n \ln(n)}/5 \). Plugging \( n = 301 \) into the formula (the number of observations that contribute to the estimation procedure at the last time point) yielded a maximum weight of 20.

**Modeling the missingness probabilities.** At each censoring node, the goal is to model the conditional probability of remaining uncensored through time \( t \), given the past, \( P(A_t = 1|\bar{A}_{t-1} = 1, \bar{L}_t, A_0) \), where \( \bar{A}_{t-1} = 1 \) denotes remaining uncensored at all censoring intervention nodes from 1 through \( t - 1 \), \( \bar{L}_t \) denotes baseline covariates and all time-varying covariates through time \( t \), and \( A_0 \) is the initial treatment. Downstream effects of treatment are potential causes of right censoring. Failing to adjust for time-dependent confounding would lead to residual bias in the effect estimate. However, for these data, adjusting for the many covariates and treatment recorded from baseline through time \( t \) is complicated by the small number of censoring events at months 5, 9, and 12 (\( n_{\text{censored},t} = 14, 7, 13 \), respectively). At these time points, regressing a rare binary outcome on a large set of covariates runs the risk of overfitting the model, even for methods that incorporate regularization (Friedman, Hastie, and Tibshirani 2001). For this reason, we designed the following two approaches:

**Approach 1.** At 3 months (time \( t_1 \)), 66 subjects were censored. We used SL to model \( P(A_t = 1|A_0, L_0) \), where \( L_0 \) includes the five determinants of randomization probabilities, and an additional nine variables selected using Lasso, based on their association with the outcome. We note that the analyst has now knowledge of the outcome when pre-specifying this automated variable selection process. For the remaining time points, where censoring events were extremely rare, we only adjusted for a single covariate, the initial treatment, by modeling \( P(A_t = 1|A_0) \), for \( t > 1 \).
Approach 2. $P\left(A_1 = 1 | A_0, \tilde{l}_0 \right)$ was modeled the same as in Approach 1. At each subsequent time point, we adjusted for treatment and time-dependent covariates by creating a one-dimensional summary score, $X_t$, as follows. SL was used to model the relationships between the outcome observed among the subjects who remained uncensored at $t = 4$, conditional on variables measured through time $t - 1$. These fitted models were used to predict $X_t$ for each subject who remained uncensored through $t - 1$. Logistic regression was then used to estimate $P(A_{t,t+1} = 1 | X_t)$.

The L-TMLE incorporated these SL-based estimates of $G_t$, and used the equivalent sample size-based truncation strategy (lower bound on each cumulative product term = 0.05). SL was also used to model the conditional expectations of the outcome given the past at each time point, working backwards from $t_4$ toward $t_0$. All calls to SL specified 20-fold cross-validation.

Step 5. Sensitivity analyses and substantive conclusion: Potential violations of the identifying causal assumptions discussed in Step 3 cast doubt on the validity of a causal interpretation of the findings. A nonparametric sensitivity analysis will investigate how robust the conclusion is with respect to the magnitude of a presumed causal gap.

5.1.2. Results

**Propensity score diagnostics.** The PS distributions in each treatment group, C-statistic of 0.62, and standardized mean differences all indicate the two groups are quite comparable at baseline (Figure 5, Supplemental Appendix Table A3).

**Estimator performance.** Point estimates and 95% CIs indicate that acupuncture reduces MWHS (Table 5). The estimate from an unadjusted complete case point treatment analysis is provided for comparison. Longitudinal Approach 1 ignores time-varying covariates, while Approach 2 incorporates time-varying data. Approach 2 shifted the L-IPTW point estimate further away from the null, with little change to the SE and CI widths. Incorporating time-varying information may have had little impact on the estimates because the majority of right censoring occurred at $t = 1$, before time-varying contributors to LTFU had manifested.

The unadjusted complete case analysis produced a more extreme finding than any adjusted analysis. L-IPTW SEs were larger than L-TMLE SEs, and confidence intervals were wider. This analysis highlights the value of capturing time-varying covariates related to treatment, outcome, and the reason for study withdrawal. Instead of merely speculating on the magnitude of bias due to time-varying confounding, the longitudinal analysis provided the opportunity to evaluate its impact, at least with respect to measured potential confounders.

5.1.3. Sensitivity Analysis and Substantive Conclusion

Sensitivity analyses reported in the original paper relied on parametric multiple imputation modeling assumptions and did not address bias due to potential unmeasured confounders. Nevertheless, they were intentionally designed to provide a conservative estimate of the treatment effect. Imputing values for the 100 missing outcomes yielded an estimated ATE of $-3.91$, consistent with the paper’s substantive conclusion that acupuncture reduced MWHS.

We ran a nonparametric sensitivity analysis to evaluate the impact of an unknown causal gap on the substantive conclusion. PS diagnostics suggest the positivity assumption is met, and consistency is guaranteed in an ITT analysis. The only likely
Figure 6. Case study 2: Sensitivity analysis showing the effect on point estimates and 95% confidence interval bounds under a presumed causal gap of up to ±12. The “Adj Units” axis shows the ratio of the causal gap to the amount of adjustment due to measured variables. Study finding shown in black.

6. Discussion

TL provides a rigorous framework for learning from data, estimating causal effects, and evaluating their reliability. Following the roadmap ensures the statistical estimation problem is clearly defined and assumptions required for a causal interpretation are explicitly stated. TMLE+SL are rigorous statistical methodologies for addressing challenges inherent in analyses of studies that deviate from the ideal RCT, including those that incorporate RWD. Nonetheless, it is always preferable to minimize foreseeable challenges in the study design phase where possible, for example, choosing a highly comparable external control arm or devising strategies for mitigating treatment nonadherence. One should also plan at the outset to collect time-varying covariates, including reason for dropout. Some inherent challenges may not be avoidable, for example, high-dimensional data or rare outcome events. Nonparametric sensitivity analyses complement traditional parametric modeling approaches to evaluate the reliability of the RWE. The G-value informs expert consideration of how large potential sources of bias would have to be to undermine the conclusion drawn from the study.

Our comparison of TMLE+SL with PS-based estimators shows the wisdom of defaulting to TMLE+SL in the planning stage. When challenges fail to arise, as in the setting of Case Study 1, most methods, including TMLE+SL, will be suitable. When confounding bias and LTFU are present the simulation studies suggest TMLE and IPTCW-SL will be more successful than any PS matching variant considered. In our simulations TMLE began to out-perform IPTW as the proportion LTFU increased. In Case Study 2, a longitudinal analysis provided the opportunity to evaluate the impact of time-varying confounding with respect to measured potential confounders. L-TMLE estimates and SEs were the most stable and had the narrowest CIs. The longitudinal analysis also illustrated a useful approach to reducing dimensionality when censoring events are rare; that is, creating a univariate summary of covariates measured through time $t$.

In the point-treatment setting, PS matching was often biased and more variable than IPW or TMLE. Post-match adjustment offered clear improvement, but an inability to handle LTFU or treatment nonadherence made it the weakest of the three approaches examined. IPTW, IPTCW (and L-IPTW) point estimates were in good agreement with TMLE (L-TMLE), but typically had larger SE and wider 95% CI. The loss in precision of the IPW estimator means that a study using IP weighting would require more data than TMLE to have the same power to detect a deviation from the null. Statistical theory shows...
that asymptotically TMLE will always be more efficient than IP weighting (van der Laan and Rose 2011). A broad range of published simulation studies demonstrate robust finite sample performance of TMLEs. Even in sparse data settings more sophisticated collaborative TMLEs can notably reduce MSE and improve CI coverage (Porter et al. 2011; Ju et al. 2017; Tackney et al. 2023). Powering a study during the design phase should always be conservative, but if external or pilot study data are available the efficiency gain from using TMLE instead of IPW may be quantified in advance to support designing a smaller study.

The guidance framework for FDA’s RWE program outlines three considerations: (i) whether the RWD are fit for use, (ii) whether the trial or study design used to generate RWE can provide adequate scientific evidence to answer or help answer the regulatory question, and (iii) whether the study conduct meets FDA regulatory requirements (FDA 2018). The TL roadmap's step-by-step approach is one that meets these considerations and is in alignment with the ICH E9 (R1) estimand guidelines. Steps 0–3 promote discussing whether the planned data collection process will ensure sufficient data type and quality to identify the target estimand, for example, whether intercurrent events, outcome, key covariate information, etc. are well defined, well captured and measured with sufficient validity. Step 4 calls for pre-specification of an analytic approach that is well suited for evaluating the statistical estimand. Step 5 assesses the reliability of the study finding. The RWE is the culmination of the entire process, supported by decisions documented each step along the way.

Trust in RWE should not be extended lightly, but we have shown how, with TL, trust can be earned when warranted. Taken together, the work products—the definition of the statistical model, causal parameter and statistical estimand; discussion of plausibility of identifying assumptions; resulting point estimates, SE, p-values, and CI; PS diagnostics; and sensitivity analyses—paint a full picture. They provide a rational basis for helping to decide whether to either accept the substantive conclusion drawn from the study findings, or to postpone acceptance until alternative information upon which to base a firm conclusion can be obtained.

**Appendix**

Data for each of the 1000 datasets generated for each plasmodesimulation study were generated as follows. \( N = 19,435 \) observations were sampled with replacement from the International Stroke Trial analytic dataset. A propensity score (PS) was generated as a complex function of the covariates. Define \( Z_1 = -2 + 2.35X_F + 0.25X_F \times 0.25X_F, Z_2 = AGE/RSBP, Z_3 = \log(RDELAY/10), \) and \( Z_4 = \log(AGE) \). The PS is given by \( P(A = 1|W) = \expit(-1.8 - 0.5Z_1 + Z_3/2 + 0.4Z_4) \). The coefficients in the PS model were chosen so that approximately 15% of observations were in the treatment group, and to introduce confounding bias. Treatment, \( A \), was generated as a Bernoulli random variable having probability \( = PS \) of being set to 1.

A binary outcome, \( Y \), was generated conditional on treatment, \( A \), and a subset of covariates \( W \) consisting of 26 main term covariates and the same four transformed versions of baseline covariates, \( Z_1, Z_2, Z_3, Z_4, P(Y = 1|A, W) = \expit(\beta_0 + \beta_1A + \beta_2AZ_\beta + \beta_2W) \). To make the relationships in the simulated data similar to those in the original dataset the values for the 30 coefficients, \( \beta \), were set equal to the estimated coefficients in a logistic regression of the true mortality outcome on the main term covariates in \( W \) (Supplemental Appendix Table A3). \( \beta_0, \beta_1, \) and \( \beta_2 \) were chosen so that the outcome occurred in approximately 17% of observations, and the treatment effect was heterogeneous. For these data the true ATE is given by \( \psi_{ATE} = -0.1381 \).

After each dataset was generated, we generated missingness indicators for studies I, II, and III as random Bernoulli variables with probability of being assigned the value 1 as follows,

I. \( P(\Delta_1 = 1|A, W) = \expit(0.9 + 1.15(0.8^{STYPE/LACS} + 0.3RDEF5_N + 0.34Z_4 + 0.2A)) \),

II. \( P(\Delta_II = 1|A, W) = \expit(-0.065 + 0.8^{STYPE/LACS} + 0.3RDEF5_N + 0.34Z_4 + 0.2A)) \),

III. \( P(\Delta_III = 1|A, W) = \expit(-0.85 + 0.17(0.8^{STYPE/LACS} + 0.3RDEF5_N + 0.34Z_4 + 0.2A)) \),

and then set \( Y_{study,j} \) to missing when \( \Delta_{study,j} = 0 \).

**Supplementary Materials**

Appendix Tables A1–A4 and sample data analysis code are provided as supplementary materials.

**Acknowledgments**

The content is the view of the author(s), and does not necessarily represent the official views of, nor an endorsement by, the FDA/IHHS or the U.S. Government.

**Disclosure Statement**

The authors report there are no competing interests to declare.

**Funding**

The authors gratefully acknowledge funding from the United States Food and Drug Administration (US FDA) pursuant to Contract 75F40119C10155.

**ORCID**

Susan Gruber https://orcid.org/0000-0003-0127-0099

**References**

Balzer, L. B., Zheng, W., van der Laan, M. J., and Petersen, M. L. (2019), “A New Approach to Hierarchical Data Analysis: Targeted Maximum Likelihood Estimation for the Causal Effect of a Cluster-Level Exposure,” *Statistical Methods in Medical Research*, 28, 1761–1780. DOI:10.1177/0962280218774936. [12]

Bang, H., and Robins, J. M. (2005), ”Doubly Robust Estimation in Missing Data and Causal Inference Models,” *Biometrics*, 61, 962–973. DOI:10.1111/j.1541-4040.2005.00377.x. [20]

Carrell, D., Gruber, S., Floyd, J. S., Bann, M. A., Cushing-Haagen, K. L., Johnson, R. L., Graham, V., Cronkite, D. J., Hazlehurt, B. L., Felcher, A. H., Bejin, C. A., Kennedy, A., Shinde, M., Karami, S., Mi, Y., Stojanovic, D., Zhao, Y., Ball, R., and Nelson, J. (in press), ”Improving Methods of Identifying Anaphylaxis for Medical Product Safety Surveillance Using Natural Language Processing and Machine Learning,” *American Journal of Epidemiology*.

Chipman, H., George, E. I., and McCulloch, R. E. (2010), ”BART: Bayesian Additive Regression Trees,” *The Annals of Applied Statistics*, 4, 266–298. DOI:10.1214/09-AOAS285. [14]

Cole, S. R., and Hernán, M. A. (2008), ”Constructing Inverse Probability Weights for Marginal Structural Models,” *American Journal of Epidemiology*, 168, 656–664. DOI:10.1093/aje/kwn164. [14]
Concorato, J., Stein, P., Dal Pan, G. J., Ball, R., and Corrigan-Curay, J. (2020), "Randomized, Observational, Interventional, and Real-World—What's in a Name?" Pharmacoepidemiology and Drug Safety, 29, 1514–1517. [12]

Corrigan-Curay, J., Sacks, L., and Woodcock, J. (2018), "Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness," JAMA, 320, 867–868. [11]

Diaz, I., and van der Laan, M. J. (2013), "Sensitivity Analysis for Causal Inference under Unmeasured Confounding and Measurement Error Problems," The International Journal of Biostatistics, 9, 149–160. [13]

DIAZ, L., and van der Laan, M. J. (2013), "Sensitivity Analysis for Causal Inference under Unmeasured Confounding and Measurement Error Problems," The International Journal of Biostatistics, 9, 149–160. [13]

DIAZ, L., and van der Laan, M. J. (2013), "Sensitivity Analysis for Causal Inference under Unmeasured Confounding and Measurement Error Problems," The International Journal of Biostatistics, 9, 149–160. [13]

Diaz, I., and van der Laan, M. J. (2013), "Sensitivity Analysis for Causal Inference under Unmeasured Confounding and Measurement Error Problems," The International Journal of Biostatistics, 9, 149–160. [13]

DIAZ, L., and van der Laan, M. J. (2013), "Sensitivity Analysis for Causal Inference under Unmeasured Confounding and Measurement Error Problems," The International Journal of Biostatistics, 9, 149–160. [13]

Diaz, I., and van der Laan, M. J. (2013), "Sensitivity Analysis for Causal Inference under Unmeasured Confounding and Measurement Error Problems," The International Journal of Biostatistics, 9, 149–160. [13]

DIAZ, L., and van der Laan, M. J. (2013), "Sensitivity Analysis for Causal Inference under Unmeasured Confounding and Measurement Error Problems," The International Journal of Biostatistics, 9, 149–160. [13]

Jaeger, D. (2012), "Covariate Balancing Propensity Score. Covariate Balancing Propensity Score," Journal of the Royal Statistical Society: Series B, 76, 243–263. DOI:10.1111/j.rssb.2012. [14]

International Stroke Trial Collaborative Group. (1997), "The International Stroke Trial (IST): A Randomised Trial of Aspirin, Subcutaneous Heparin, both, or Neither among 19435 Patients with Acute Ischaemic Stroke," The Lancet, 349, 1569–1581. [14]

Ju, C., Wyss, R., Franklin, J. M., Schneewiss, S., Hägström, J., and van der Laan, M. J. (2017), "Collaborative-Controlled LASSO for Constructing Propensity Score-based Estimators in High-Dimensional Data," Statistical Methods in Medical Research, 28, 1044–1063. DOI:10.1177/0962280217744588. [23]

Kahale, L. A., Diab, B., Khamis, A. M., Chang, Y., Cruz Lopes, L., Agarwal, A., Li, L., Mustafa, R. A., Koujianian, S., Waziry, R., Busse, J. W., Dakik, A., Guyatt, G., and Aki, E. A. (2019), "Potentially Missing Data are Considerably More Frequent than Definitely Missing Data: A Methodological Survey of 638 Randomized Controlled Trials," Journal of Clinical Epidemiology, 106, 18–31. [12]

Kempker, R. R., Miklashwili, L., Zhao, Y., Benkeser, D., Barbakadze, K., Dablishwili, N., Avaliani, Z., Peloquin, C. A., Blumberg, H. M., and Kipiani, M. (2020), "Clinical Outcomes Among Patients With Drug-resistant Tuberculosis Receiving Bedaquiline- or Delamanid-Containing Regimens," Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America, 71, 2334–2344. DOI:10.1093/cid/ciz1107. [12]

Kreif, N., Tran, L., Grieve, R., De Stavola, B., Tasker, R. C., and Petersen, M. (2017), "Estimating the Comparative Effectiveness of Feeding Interventions in the Pediatric Intensive Care Unit: A Demonstration of Longitudinal Targeted Maximum Likelihood Estimation," American Journal of Epidemiology, 186, 1370–1379. DOI:10.1093/aje/kov213. [12]

Lendle, S. D., Schwab, J., Petersen, M. L., and van der Laan, M. J. (2017), "Imle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data," Journal of Statistical Software, 81, 1–21. DOI:10.18637/jss.v081.i01. [14]

Lendle, S. D., Fireman, B., and van der Laan, M. J. (2013), "Targeted Maximum Likelihood Estimation in Safety Analysis," Journal of Clinical Epidemiology, 66, S91–S98. DOI:10.1016/j.jclinepi.2013.02.017. [13]

Ling, A. Y., Montez-Rath, M. E., Mathur, M. B., Kappahn, K., and Desai, M. (2019), "How to Apply Multiple Imputation in Propensity Score Matching with Partially Observed Confounders: A Simulation Study and Practical Recommendations," arXiv Preprint arXiv:1904.07408. [14]

Pearl, J. (2009), Causality, Cambridge: Cambridge University Press. [12]

Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M., and van der Laan, M. (2014), "Targeted Maximum Likelihood Estimation for Dynamic and Static Longitudinal Marginal Structural Working Models," Journal of Causal Inference, 2, 147–185. [20]

Petersen, M. L., and van der Laan, M. J. (2014), "Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation," Epidemiology (Cambridge, Mass.), 25, 418–426. DOI:10.1097/EDE.0000000000000078. [12]

Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S., and van der Laan, M. J. (2015), "Mortality Prediction in Intensive Care Units with the Super ICU Learner Algorithm (SICULA): A Population-based Study," The Lancet Respiratory Medicine, 3, 42–52. [12]

Polley, E., LeDell, E., Kennedy, C., and van der Laan, M. (2019), SuperLearner: Super Learner Prediction. R package version 2.0-26. Available at https://CRAN.R-project.org/package=SuperLearner. [14]

Porter, K. E., Gruber, S., van der Laan, M. J., and Sekhon, J. S. (2011), "The Relative Performance of Targeted Maximum Likelihood Estimators," The International Journal of Biostatistics, 7, 1–34. [13,23]

R Core Team (2020), R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org/. [13]

Ramasahai, R. R., Grieve, R., and Sekhon, J. S. (2011), "Extending Iterative Matching Methods: An Approach to Improving Covariate Balance That Allows Prioritisation," Health Services and Outcomes Research Methodology, 11, 95–114. DOI:10.1007/s10742-011-0075-5. [14]
Rosenbaum, P. R., and Rubin, D. B. (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55. [12]

Rosenblum, M., and van der Laan, M. J. (2010), “Simple, Efficient Estimators of Treatment Effects in Randomized Trials using Generalized Linear Models to Leverage Baseline Variables,” *The International Journal of Biostatistics*, 6, 1–41. DOI:10.2202/1557-4679.1138. [15]

Sandrock, P., Niewada, M., and Czlonkowska, A. (2011), International Stroke Trial database (version 2), University of Edinburgh. Department of Clinical Neurosciences. DOI:10.7488/ds/104 [14]

Simon, G. E., Bindman, A. B., Dreyer, N. A., Platt, R., Watanabe, J. H., Horberg, M., Hernandez, A., and Califf, R. M. (2022), “When Can We Trust Real-World Data To Evaluate New Medical Treatments?,” *Clinical Pharmacology and Therapeutics*, 111, 24–29. DOI:10.1002/cpt.2252. [11]

Stuart, E. A. (2010), ”Matching Methods for Causal Inference: A Review and a Look Forward,” *Statistical Science*, 25, 1–21. DOI:10.1214/09-STS313. [14]

Tackney, M. S., Morris, T., White, I., Leyrat, C., Diaz-Ordaz, K., and Williamson, E. (2023), “A Comparison of Covariate Adjustment Approaches Under Model Misspecification In Individually Randomized Trials,” *Trials*, 24, 1–8. DOI:10.21203/rs.3.rs-1053600/v1. [23]

van der Laan, M. J., and Robins, J. (2003), *Unified Methods for Censored Longitudinal Data and Causality*, New York: Springer. [20]

van der Laan, M. J., and Rose, S. (2011), *Targeted Learning: Causal Inference for Observational and Experimental Data*, New York: Springer. [11,13,23]

VanderWeele, T., and Ding, P. (2017), “Sensitivity Analysis in Observational Research: Introducing the E-value,” *Annals of Internal Medicine*, 167, 268–274. [13]

Vickers, A. J., and Czlonkowska, A. (2011), International Stroke Trial database (version 2), University of Edinburgh. Department of Clinical Neurosciences. DOI:10.7488/ds/104 [14]

Simon, G. E., Bindman, A. B., Dreyer, N. A., Platt, R., Watanabe, J. H., Horberg, M., Hernandez, A., and Califf, R. M. (2022), “When Can We Trust Real-World Data To Evaluate New Medical Treatments?,” *Clinical Pharmacology and Therapeutics*, 111, 24–29. DOI:10.1002/cpt.2252. [11]

Stuart, E. A. (2010), ”Matching Methods for Causal Inference: A Review and a Look Forward,” *Statistical Science*, 25, 1–21. DOI:10.1214/09-STS313. [14]

Tackney, M. S., Morris, T., White, I., Leyrat, C., Diaz-Ordaz, K., and Williamson, E. (2023), “A Comparison of Covariate Adjustment Approaches Under Model Misspecification In Individually Randomized Trials,” *Trials*, 24, 1–8. DOI:10.21203/rs.3.rs-1053600/v1. [23]

van der Laan, M. J., and Robins, J. (2003), *Unified Methods for Censored Longitudinal Data and Causality*, New York: Springer. [20]

van der Laan, M. J., and Rose, S. (2011), *Targeted Learning: Causal Inference for Observational and Experimental Data*, New York: Springer. [11,13,23]

VanderWeele, T., and Ding, P. (2017), “Sensitivity Analysis in Observational Research: Introducing the E-value,” *Annals of Internal Medicine*, 167, 268–274. [13]

Vickers, A. J., and Czlonkowska, A. (2011), International Stroke Trial database (version 2), University of Edinburgh. Department of Clinical Neurosciences. DOI:10.7488/ds/104 [14]

Simon, G. E., Bindman, A. B., Dreyer, N. A., Platt, R., Watanabe, J. H., Horberg, M., Hernandez, A., and Califf, R. M. (2022), “When Can We Trust Real-World Data To Evaluate New Medical Treatments?,” *Clinical Pharmacology and Therapeutics*, 111, 24–29. DOI:10.1002/cpt.2252. [11]

Stuart, E. A. (2010), ”Matching Methods for Causal Inference: A Review and a Look Forward,” *Statistical Science*, 25, 1–21. DOI:10.1214/09-STS313. [14]