Development and Application of the Experiment-Selector Cross-Validated Targeted Maximum Likelihood Estimator

By

Lauren Elizabeth Eyler Dang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark van der Laan, Chair
Professor Maya Petersen
Professor Alan Hubbard
Professor Ziad Obermeyer

Spring 2023

Development and Application of the Experiment-Selector Cross-Validated Targeted Maximum Likelihood Estimator

Abstract

Development and Application of the Experiment-Selector Cross-Validated Targeted Maximum Likelihood Estimator

by

Lauren Elizabeth Eyler Dang

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Mark van der Laan, Chair

This dissertation encompasses the development and application of the experiment-selector cross-validated targeted maximum likelihood estimator (ES-CVTMLE) for analyzing hybrid randomized-external data studies. The goal of these hybrid designs is to augment a small randomized controlled trial (RCT) with external data – in the form of the control arm(s) of previous trials or real-world healthcare data (RWD) – in order to increase power. Of course, inclusion of RWD may also increase the causal gap, defined as the difference between the causal effect of interest and the statistical parameter that we will estimate from the data. The primary statistical challenges are 1) excluding external data that would introduce bias of a magnitude large enough to worsen coverage for the causal effect while still including unbiased external data frequently enough to improve power and 2) constructing confidence intervals that appropriately reflect that the causal gap may not be zero when external data are integrated.

In Chapter 1, we describe the development of the ES-CVTMLE methodology, focusing on the case where only external controls are available. We consider two methods of estimating the causal gap: 1) a function of the difference in conditional mean outcome under control between the RCT and combined experiments and 2) the estimated average treatment effect on a negative control outcome. We then define criteria for selecting the experiment (RCT alone or RCT combined with external data) that optimizes the estimated bias-variance tradeoff. To separate the data used for experiment selection from the data used for effect estimation, we develop an experiment-selector cross-validated targeted maximum likelihood estimator. We define the asymptotic distribution of the ES-CVTMLE under varying magnitudes of bias and construct confidence intervals by Monte Carlo simulation. We demonstrate the performance of the ES-CVTMLE compared to three other estimators for hybrid randomized-external data designs using simulations and a re-analysis of the LEADER trial of the effect of liraglutide versus placebo on cardiovascular outcomes.

In Chapter 2, we describe the development of the EScvtmle R software package to implement the method described in Chapter 1. The software package also extends this methodology to allow for integration of external data participants with both the active treatment and control arms of the trial. We include vignettes demonstrating use of the EScvtmle package with the publicly available WASH Benefits Bangladesh cluster RCT dataset.

The real data examples in Chapters 1 and 2 rely on following the Roadmap for Causal and Statistical Inference, a structured process that guides the design, analysis, and interpretation of

studies anywhere on the spectrum from a traditional RCT to a fully observational study. In Chapter 3, we describe this Causal Roadmap to an audience of clinical and translational researchers. We also extend the Roadmap framework to consider how outcome-blind simulations may be used for quantitative comparison of the characteristics of different potential study designs.

Chapter 4 represents the culmination of the previous work; we use a case study of semaglutide and cardiovascular outcomes to demonstrate application of this extended version of the Causal Roadmap to compare study designs involving traditional RCTs with a hybrid randomized-external data design. We demonstrate how following the Causal Roadmap can help to define an external control arm in a way that improves the plausibility of causal identification assumptions. We then use simulations to demonstrate the tradeoffs between each of these potential designs. Finally, we present a real data analysis using the ES-CVTMLE to estimate the effect of oral semaglutide versus standard-of-care on risk of major adverse cardiovascular events based on the PIONEER 6 RCT and considering augmentation with RWD from Optum's de-identified Clinformatics® Data Mart Database (CDM) (2007-2022).

To Eric, Dad, Mom, Lindsay, and Ahsoka. I am so thankful for all of you and for all of your support!

Acknowledgements

I want to express my deepest thanks to the following amazing people who helped and supported me during this PhD:

- Mark van der Laan: Thank you for taking a chance on a trainee with a very limited math background, patiently explaining complex concepts to me over many zoom hours, and supporting me in choosing my own adventure.
- Maya Petersen: Thank you for being an outstanding mentor and role model, giving me so many opportunities and supporting me through each of them, and being unflaggingly dedicated to the causes you believe in no matter how busy or difficult life gets.
- Alan Hubbard: Thank you for bringing me along on this biostatistical adventure in the first place and for always being willing to talk through ideas despite my many stumbles and detours along the way.
- Ziad Obermeyer: Thank you for serving on my dissertation committee! I greatly appreciate your time.
- To my Novo Nordisk and UNC co-authors: Edwin Fong, Jens Tarp, Kim Clemmensen, Trine Abrahamsen, Henrik Ravn, Kajsa Kvist, and John Buse. I have really enjoyed working with all of you these past few years. Edwin, Jens, and Kim, thank you for conducting the very complicated real data analysis that appears in Chapter 4 of this dissertation. Jens, thank you also for your help processing the LEADER dataset for the analysis that appears in Chapter 1. I really appreciate your dedication to these projects and your contributions to this dissertation and our papers!
- To the Forum for the Integration of Observational and Randomized Data (FIORD) Workshop participants (including Susan Gruber, Hana Lee, Issa Dahabreh, Liz Stuart, Brian Williamson, Richie Wyss, Ivan Diaz, Debashis Ghosh, Emre Kiciman, Demissie Alemayehu, Martin Ho, Kat Hoffman, Carla Vossen, Ray Huml, Henrik Ravn, Kajsa Kvist, Richard Pratley, Mei-Chiung Shih, Gene Pennello, David Martin, Salina Waddy, Charlie Barr, Mouna Akacha, John Concato, and John Buse): Thank you for your many comments and suggestions that helped to improve Chapter 3 of this dissertation.
- To my family: Thank you for all your love and support as I wander down this winding career path of mine. I could not have done it without you.

Contents

1 The Experiment-Selector CV-TMLE

1.1 Introduction

With the growing availability of observational data from sources such as registries, electronic health records, or the control arms of previous trials, the power of randomized controlled trials (RCTs) could potentially be improved while randomizing fewer participants to control status if we were able to incorporate real-world data (RWD) in the analysis [1, 2, 3]. Running an adequately-powered trial without external control data may be infeasible for rare diseases [4]. For severe diseases without effective treatments or pediatric approvals of medications that have been shown to be safe and efficacious in adults, inclusion of external control data may allow more trial participants to be randomized to receive a potentially beneficial medication instead of placebo [2, 5].

Yet combining these data types comes with the risk of introducing bias from multiple sources, including measurement error, selection bias, and confounding [6]. Data fusion estimators, discussed in detail in Related Literature below, aim to estimate the bias that may be introduced by incorporating real-world data in order to decide whether to include RWD or how to weight RWD in a hybrid RCT-observational analysis. These estimators may take a Bayesian [1, 7, 8, 3] or a frequentist [2, 9, 10, 11, 12, 13] approach and use different criteria for inclusion of RWD and different methods of confidence interval construction.

A key insight from this literature is that there is an inherent tradeoff between maximizing power when unbiased RWD are available and maintaining close to nominal coverage across the spectrum of potential magnitudes of RWD bias [11, 13]. The strengths and limitations of existing methods led us to consider an alternate approach to augmenting the control arm of an RCT with external data that incorporates multiple estimates of bias to boost potential power gains while providing robust inference despite violations of necessary identification assumptions. Framing the decision of whether to integrate RWD (and by extension, which RWD to integrate) as a problem of data-adaptive experiment selection, we develop a novel cross-validated targeted maximum likelihood estimator for this context that 1) incorporates an estimate of the average treatment effect on a negative control outcome (NCO) into the bias estimate, 2) uses cross-validation to separate bias estimation from effect estimation, and 3) constructs confidence intervals by sampling from the estimated limit distribution of this estimator, where the sampling process includes an estimate of the bias, further promoting accurate inference.

The remainder of this chapter is organized as follows. In Section 2, we discuss related data fusion estimators. In Section 3, we introduce the problem of data-adaptive experiment selection and discuss issues of causal identification, including estimation of bias due to inclusion of RWD. In Section 4, we introduce potential criteria for including RWD based on optimizing the bias-variance tradeoff and utilizing the estimated effect of treatment on an NCO. In Section 5, we develop an extension of the cross-validated targeted maximum likelihood estimator (CV-TMLE) [14, 15] for this

new context of data-adaptive experiment selection and define the limit distribution of this estimator under varying amounts of bias. In Section 6, we set up a simulation to assess the performance of our estimator and describe four potential comparator methods: two test-then-pool approaches [2], one method of Bayesian dynamic borrowing [3], and a difference-in-differences (DID) approach to adjusting for bias based on a negative control outcome [16, 17]. We also introduce a CV-TMLE based version of this DID method. In Section 7, we compare the causal coverage, power, bias, variance, and mean squared error of the experiment-selector CV-TMLE to these four methods as well as to a CV-TMLE and t-test for the RCT only. In Section 8, we demonstrate the use of the experiment-selector CV-TMLE to distinguish biased from unbiased external controls in a real data analysis of the effect of liraglutide versus placebo on improvement in glycemic control in the Central/South America subgroup of the LEADER trial.

1.2 Related literature

A growing literature highlights different strategies for combined RCT-RWD analyses. One set of approaches, known as Bayesian dynamic borrowing, generates a prior distribution of the RCT control parameter based on external control data, with different approaches to down-weighting the observational information [1, 7, 8, 3]. These methods generally require assumptions on the distributions of the involved parameters, which may significantly impact the effect estimates [18, 5]. While these methods can decrease bias compared to pooling alone, multiple studies have noted either increased type 1 error or decreased power when there is heterogeneity between the historical and RCT control groups [5, 2, 18, 19, 20].

This tradeoff between the ability to increase power with unbiased external data and the ability to control type 1 error across all potential magnitudes of bias has also been noted in the frequentist literature [11, 13]. A simple "test-then-pool" strategy for combining RCT and RWD, described by Viele et al.[2], involves a hypothesis test that the mean outcomes are equal in the RCT and RWD control arms; datasets are only combined if the null hypothesis of the test is not rejected. However, when the RCT is small, tests for inclusion of RWD are also underpowered, and so observational controls may be inappropriately included even when the test's null hypothesis is not rejected [21]. Thus, such approaches are subject to inflated type 1 error in exactly the settings in which inclusion of external controls is of greatest interest.

Subsequently, several estimators that are more conservative in their ability to maintain nominal type 1 error control have been proposed. For example, Rosenman et al.[9] have built on the work of Green and Strawderman[22] in adapting the James-Stein shrinkage estimator [23] to weight RCT and RWD effect estimates in order to estimate stratum-specific average treatment effects. Another set of methods aims to minimize the mean squared error of a combined RCT-RWD estimator, with various criteria for including RWD or for defining optimal weighted combinations of RCT and RWD [10, 11, 12, 13]. These studies reveal the challenge of optimizing the bias-variance tradeoff when bias must be estimated. Oberst et al.[13] note that estimators that decrease variance most with unbiased RWD also tend to have the largest increase in relative mean squared error compared to the RCT only when biased RWD is considered. Similarly, Chen et al.[11] show that if the magnitude of bias introduced by incorporating RWD is unknown, the optimal minimax confidence interval length for their anchored thresholding estimator is achieved by an RCT-only estimator, again demonstrating that both power gains and guaranteed type I error control should not be expected. Yang et al.[10], Chen et al.[11], and Cheng and Cai[12] introduce tuning parameters for their estimators to modify

this balance. Because no estimator is likely to outperform all others both by maximizing power and maintaining appropriate type 1 error in all settings, different estimators may be beneficial in different contexts where one or the other of these factors is a greater priority. While these methods focus on estimating either the conditional average treatment effect [10, 12] or the average treatment effect [11, 13] in contexts when treatment is available in the external data, in this chapter we focus on the setting where a medication has yet to be approved in the real world.

An alternate approach to estimating bias, used mostly for observational data analyses, involves the use of an NCO. Because the treatment does not affect an NCO, evidence of an association between the treatment and this outcome is indicative of bias [24]. Authors including Sofer et al.[16], Shi et al.[25], and Miao et al.[26] have developed methods of bias adjustment using an NCO. Yet because there may be unmeasured factors that confound the relationship between the treatment and the true outcome that do not confound the relationship between the treatment and the NCO, an NCO-based bias estimate near zero does not rule out residual bias [24].

In summary, methods that estimate bias to evaluate whether to include RWD or how to weight RWD in a combined analysis most commonly rely either on a comparison of mean outcomes or effect estimates between RCT and RWD (e.g., [2, 3, 10, 13]) or on the estimated average treatment effect on an NCO (e.g., [17]). The latter approach requires additional assumptions regarding the quality of the NCO [17, 24]. Bias estimation is a challenge for both approaches, leading to a tradeoff between the probability that information from unbiased RWD is included and the probability that information from biased RWD is excluded [11, 13]. We discuss both options for bias estimation and our proposal to combine information from both sources below.

1.3 Causal Roadmap for hybrid RCT-RWD trials

In this section, we follow the causal inference roadmap described by Petersen and van der Laan[27] to explain this data fusion challenge. Please refer to Supplementary Table A.1 in Appendix A.1 for a list of symbols used in this chapter. For a hybrid RCT-RWD study, let S indicate the experiment being analyzed, where $s_i = 0$ indicates that individual i participated in an RCT, $s_i \in \{1, ..., K\}$ indicates that individual i participated in one of K potential observational cohorts, and $S \in \{0, s\}$ indicates an experiment combining an RCT with dataset s. We have a binary intervention, A, a set of baseline covariates, W, and an outcome Y. W may affect inclusion in the RCT versus RWD. Assignment to active treatment, A, is randomized with probability p for those in the RCT and set to 0 (standard of care) for those in the RWD, because the treatment has yet to be approved. Thus, A is only affected by S and p, not directly by W or any exogenous error. Y may be affected by W, A, and potentially also directly by S. The unmeasured exogenous errors $U = (U_W, U_S, U_Y)$ for each of these variables could potentially be dependent. The full data then consist of both endogenous and exogenous variables. Our observed data are n independent and identically distributed observations $O_i = (W_i, S_i, A_i, Y_i)$ with true distribution P_0.

A common causal target parameter for RCTs is the average treatment effect (ATE). With multiple available datasets, there are multiple possible experiments we could run to evaluate the ATE for the population represented by that experiment, where each experiment includes S=0 with or without external control dataset s. With counterfactual outcomes [28] defined as the outcome an individual would have had if they had received treatment (Y^1) or standard of care (Y^0), there are thus multiple potential causal parameters that we could target, one for each potential experiment:
$\Psi_s^F(P_{U,O}) = E_{W|S\in\{0,s\}}[E(Y^1 - Y^0|W, S \in \{0, s\})]$ for $s \in \{0, ..., K\}$.

1.3.1 Identification

Next, we discuss whether each of the potential causal parameters, $\Psi_s^F(P_{U,O})$, is identifiable from the observed data.

Lemma 1: For each experiment with $S \in \{0, s\}$, under **Assumptions 1 and 2a-b** below, the causal ATE, $\Psi_s^F(P_{U,O})$, is identifiable from the observed data by the g-computation formula [29], with statistical estimand

$$\Psi_s(P_0) = E_{W|S\in\{0,s\}}[E_0[Y|A=1, S\in\{0,s\}, W] - E_0[Y|A=0, S\in\{0,s\}, W]]. \quad (1.1)$$

Assumption 1 (Positivity (e.g., [30, 31])): $P(A=a|W=w, S\in\{0,s\}) > 0$ *for all* $a \in A$ *and all* w *for which* $P(W=w, S\in\{0,s\}) > 0$. This assumption is true in the RCT by design and may be satisfied for other experiments by removing RWD controls whose W covariates do not have support in the trial population.

Assumption 2 (Mean Exchangeability (e.g., [32, 33])):
As described by Rudolph et al.[32] and subsequently named by Dahabreh et al.[33],
Assumption 2a ("Mean exchangeability in the trial" [33]): $E[Y^a|W, S=0, A=a] = E[Y^a|W, S=0]$. This assumption is also true by the design of the RCT.
Assumption 2b ("Mean exchangeability over S" [33]): $E[Y^a|W, S=0] = E[Y^a|W, S\in\{0,s\}]$ *for every* $a \in A$. **Assumption 2b** may be violated if unmeasured factors affect trial inclusion or if being in the RCT directly affects adherence or outcomes [32, 34]. Dahabreh et al.[34] note that **Assumption 2b** is more likely to be true for pragmatic RCTs integrated with RWD from the same healthcare system. Nonetheless, we may not be certain whether **Assumption 2b** is violated in practice.

1.3.2 Bias estimation

One approach to concerns about violations of **Assumption 2b** would be to target a causal parameter that we know is identifiable from the observed data. As noted by Hartman et al.[35], Balzer et al.[36], and Dahabreh et al.[37], we may consider interventions not only on treatment assignment but also on trial participation. The difference in the outcomes an individual would have had if they had received active treatment and been in the RCT ($Y^{a=1,s=0}$) compared to if they had received standard of care and been in the RCT ($Y^{a=0,s=0}$), averaged over a distribution of covariates that are represented in the trial, gives a causal ATE of A on Y in the population defined by that experiment. Under **Assumptions 1 and 2a**, this "ATE-RCT" parameter for any experiment, $\tilde{\Psi}_s^F(P_{U,O}) = E_{W|S\in\{0,s\}}[E(Y^{a=1,s=0} - Y^{a=0,s=0}|W, S\in\{0,s\})]$, is equal to the following statistical estimand:

$$\tilde{\Psi}_s(P_0) = E_{W|S\in\{0,s\}}[E_0[Y|A=1, S=0, W] - E_0[Y|A=0, S=0, W]]. \quad (1.2)$$

Nonetheless, by estimating this parameter, we would not gain efficiency compared to estimating the sample average treatment effect for the RCT only [38].

Another general approach to addressing concerns regarding violations of **Assumption 2b** would be to estimate the causal gap or bias due to inclusion of external controls. In order to further explore this option, we consider two causal gaps as the difference between one of our two potential causal parameters and the statistical estimand $\Psi_s(P_0)$ for a given experiment with $S \in \{0, s\}$:

1. **Causal Gap 1:** $\Psi_s^F(P_{U,O}) - \Psi_s(P_0)$
2. **Causal Gap 2:** $\tilde{\Psi}_s^F(P_{U,O}) - \Psi_s(P_0)$

While these causal gaps are functions of the full and observed data, we can estimate a statistical gap that is only a function of the observed data as

$$\begin{aligned} \Psi_s^{\#}(P_0) &= \Psi_s(P_0) - \tilde{\Psi}_s(P_0) \\ = E_{W|S\in\{0,s\}}[E_0[Y|A=0,S=0,W]] &- E_{W|S\in\{0,s\}}[E_0[Y|A=0,S\in\{0,s\},W]] \end{aligned} \tag{1.3}$$

Lemma 2: Causal and Statistical Gaps for an experiment with $S \in \{0, s\}$
*If **Assumption 2b** is true, then* $\Psi_s^F(P_{U,O}) = \tilde{\Psi}_s^F(P_{U,O})$, $\Psi_s^{\#}(P_0) = 0$

Causal Gap 1: $\Psi_s^F(P_{U,O}) - \Psi_s(P_0) = 0$
Causal Gap 2: $\tilde{\Psi}_s^F(P_{U,O}) - \Psi_s(P_0) = 0$

$\Psi_s^{\#}(P_0)$ may thus be used as evidence of whether **Assumption 2b** is violated. If we were to bias correct our estimate $\Psi_s(P_0)$ by subtracting $\Psi_s^{\#}(P_0)$, we would again be estimating $\tilde{\Psi}_s(P_0)$, with no gain in efficiency compared to estimating the sample ATE from the RCT only [38]. Nonetheless, the information from estimating $\Psi_s^{\#}(P_0)$ may still be incorporated into an experiment selector, $s_n^\star$, discussed below.

1.4 Potential experiment selection criteria

A natural goal for experiment selection would be to optimize the bias-variance tradeoff for estimating a causal ATE. Such an approach of determining combinations of RCT and RWD that minimize the estimated mean squared error is taken by Yang et al.[10], Cheng and Cai[12], Chen et al.[11], and Oberst et al.[13]. Next, we discuss the challenge of selecting a truly optimal experiment when bias must be estimated from the data. We then introduce a novel experiment selector that incorporates bias estimates based on both the primary outcome and a negative control outcome.

Ideally, we would like to construct a selector that is equivalent to the oracle selector of the experiment that optimizes the bias-variance tradeoff for our target parameter:

$$s_0 = \underset{s}{argmin} \frac{\sigma^2_{D^*_{\Psi_s}}}{n} + (\Psi_s^{\#}(P_0))^2$$

where

$$\begin{aligned} &D^*_{\Psi_s}(O) = \\ &\frac{I(S\in\{0,s\})}{P(S\in\{0,s\})}\left(\left(\frac{I(A=1)}{g_0^a(A=1|W,S\in\{0,s\})} - \frac{I(A=0)}{g_0^a(A=0|W,S\in\{0,s\})}\right)(Y - Q_0^{\{0,s\}}(S\in\{0,s\},A,W))\right. \\ &\left. + Q_0^{\{0,s\}}(S\in\{0,s\},1,W) - Q_0^{\{0,s\}}(S\in\{0,s\},0,W) - \Psi_s(P_0)\right) \end{aligned}$$

is the efficient influence curve of $\Psi_s(P_0)$, $Q_0^{\{0,s\}}(S\in\{0,s\},A,W) = E_0[Y|S\in\{0,s\},A,W]$, and $g_0^a(A=a|W,S\in\{0,s\}) = P_0(A=a|W,S\in\{0,s\})$. Our statistical estimand of interest is then $\Psi_{s_0}(P_0)$.

The primary challenge is that s_0 must be estimated. We thus define an empirical bias squared plus variance (**"b2v"**) selector,

$$s_n^\star = \underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + (\hat{\Psi}_s^{\#}(P_n))^2 \tag{1.4}$$

If, for a given experiment with $S \in \{0, s\}$, $\Psi_s^{\#}(P_0)$ were given and small relative to the standard error of the ATE estimator for that experiment, nominal coverage would be expected for the causal target parameter. If bias were large relative to the standard error of the ATE estimator for the

RCT, then the RWD would be rejected, and only the RCT would be analyzed. One threat to valid inference using this experiment selection criterion is the case where bias is of the same order as the standard error $\sigma_{D^*_{\Psi_s}}/\sqrt{n}$, risking decreased coverage. We could require a smaller magnitude of bias by putting a penalty term in the denominator of the variance as $s^\star_n = \underset{s}{argmin}\ \hat{\sigma}^2_{D^*_{\Psi_s}}/(n * c(n)) + (\hat{\Psi}^{\#}_s(P_n))^2$ where $c(n)$ is either a constant or some function of n. A similar approach is taken by Cheng and Cai[12] who multiply the bias term by a penalty and determine optimal weights for RCT and RWD estimators via L1-penalized regression. However, finite sample variability may lead to overestimation of bias for unbiased RWD and underestimation of bias similar in magnitude to $\sigma_{D^*_{\Psi_s}}/\sqrt{n}$. In order to make $c(n)$ large enough to prevent selecting RWD that would introduce bias of a magnitude that could decrease coverage for the causal parameter, we would also prevent unbiased RWD from being included in a large proportion of samples.

This challenge exists for any method that bases inclusion of RWD on differences in the mean or conditional mean outcome under control for a small RCT control arm versus a RWD population. It also suggests that having additional knowledge beyond this information may help the selector distinguish between RWD that would introduce varying degrees of bias. Intuitively, if we are not willing to assume mean exchangeability, information available in the RCT alone is insufficient to estimate bias from including real world data in the analysis precisely enough to guarantee inclusion of extra unbiased controls and exclusion of additional controls that could bias the effect estimate; if the RCT contained this precise information about bias, we would be able to estimate the ATE of A on Y from the RCT precisely enough to not require the real world data at all. Conversely, if we were willing to assume mean exchangeability, then simply pooling RCT and RWD would provide optimal power gains but also fully relinquish the protection to inference afforded by randomization.

1.4.1 Additional knowledge to improve experiment selector

One additional source of information regarding bias is the estimated effect of the treatment on a negative control outcome. An NCO is not affected by the treatment but is affected by unmeasured factors that are associated with both the treatment and the outcome [24]. A non-zero estimated ATE of treatment on the NCO is therefore either due to finite sample variability or due to these unmeasured common causes. NCOs have been used primarily in observational analyses to detect and/or adjust for unmeasured confounding [16, 26, 25, 17]. In order to fully adjust for bias using an NCO, we must assume U-comparability: that the unmeasured factors that confound the treatment-outcome relationship are the same as the unmeasured factors that confound the treatment-NCO relationship [24]. When this assumption is not met, the estimated effect of treatment on the NCO represents some unknown percentage of the total bias that comes from incorporating real-world data.

Again using the g-computation formula [29], we may define an estimand of the ATE of treatment on the NCO as

$$\Phi_s(P_0) = E_{W|S\in\{0,s\}}[E_0[NCO|W, A = 1, S \in \{0, s\}] - E_0[NCO|W, A = 0, S \in \{0, s\}]] \quad (1.5)$$

Then, we could add our estimate $\hat{\Phi}_s(P_n)$ to our estimate of the bias, with selector "**+nco**":

$$s^{\star\star}_n = \underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + (\hat{\Psi}^{\#}_s(P_n) + \hat{\Phi}_s(P_n))^2 \quad (1.6)$$

Because $\Psi^{\#}_{s=0}(P_0)$ is deterministically 0 with only RCT data, whereas $\hat{\Phi}_{s=0}(P_n)$ may be estimated but with greater variability than $\hat{\Phi}_{s>0}(P_n)$ due to the smaller size of the RCT compared to

RCT plus RWD, $s_n^{\star\star}$ helps to promote the inclusion of unbiased external controls. If only biased external controls are available, however, $\Phi_{s>0}(P_0)$ has a larger magnitude for the combined RCT-biased RWD experiment because unmeasured confounding makes this statistical quantity not truly zero. We would expect that including $\hat{\Phi}_s(P_n)$ in the selector should thereby increase the probability that biased RWD is rejected. Yet we must note that this selector relies on the assumption that the unmeasured factors that affect the treatment-outcome relationship affect the treatment-NCO relationship in the same direction. Otherwise the sum $\hat{\Psi}_s^{\#}(P_n) + \hat{\Phi}_s(P_n)$ will negate some of the true bias from including external controls. This consideration should be made when selecting an appropriate NCO, but is still weaker than assumptions necessary for bias adjustment using an NCO.

We also consider selector "**nco only**" based only on $\hat{\Phi}_s(P_n)$:

$$s_n^{\star\star\star} = \underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + (\hat{\Phi}_s(P_n))^2 \tag{1.7}$$

Nonetheless, because we cannot learn from the data what percentage of the true bias is accounted for by this estimate, we choose to combine rather than replace our estimate $\hat{\Psi}_s^{\#}(P_n)$ with this information. We will compare these options with the originally-proposed selector $s_n^{\star}$. The advantage of $s_n^{\star\star}$ compared to introducing a penalty term in the denominator of the selector's variance term is that the penalty term makes the selector less likely to include *any* real-world data, while $s_n^{\star\star}$ has the potential to promote inclusion of unbiased RWD while discouraging the inclusion of biased RWD.

1.5 CV-TMLE for data-adaptive experiment selection

Now that we have defined potential experiment-selection criteria, we must use the data both to select and analyze the optimal experiment. If we select $s_n^{\star}$ in a manner that is not outcome-blind, we should not expect to obtain valid inference if we both select the experiment and evaluate our target parameter based on the same data [15]. Cross-validated targeted maximum likelihood estimation (CV-TMLE) was previously developed as a method to obtain valid inference for other data-adaptive target parameters [14, 15, 39]. We build on this previous work by developing a CV-TMLE for data-adaptive experiment selection, which poses new challenges for inference, described below.

First, we randomly split the data into V samples with an experiment-selection set consisting of $(V-1)/V$ of the data and an estimation set consisting of $1/V$. For each split, v, the estimation set has empirical distribution $P_{n,v}$ with estimation set subjects assigned $\bar{V}_i = v$. The experiment-selection set has empirical distribution P_{n,v^c}, and therefore the experiment-selection observations have $\bar{V}_i \neq v$. For each split, the experiment-selection set is used to define a data-adaptive target parameter mapping based on a fold-specific selection criterion, $s_n^{\star}(v^c)$. The fold-specific target parameter then becomes $\Psi^F_{s_n^{\star}(v^c)}(P_{U,O})$, the causal ATE of A on Y in the experiment selected based on the experiment-selection set for fold v. The overall target parameter, ψ_0, and statistical estimand, $\psi_{n,0}$, are then averages of the fold-specific parameters and estimands:

$$\psi_0 = \tfrac{1}{V} \textstyle\sum_{v=1}^{V} \Psi^F_{s_n^{\star}(v^c)}(P_{U,O})$$
$$\psi_{n,0} = \tfrac{1}{V} \textstyle\sum_{v=1}^{V} \Psi_{s_n^{\star}(v^c)}(P_0)$$

Our modified ES-CVTMLE estimator for data-adaptive experiment-selection is then:

$$\psi_n = \frac{1}{V} \sum_{v=1}^{V} \hat{\Psi}_{s_n^{\star}(v^c)}(Q_{n,v}^{\{0,s\},*})$$

where $Q_{n,v}^{\{0,s\},*}$ indicates training of initial estimators of the outcome regression $Q_{n,v^c}^{\{0,s\}}$ and treatment mechanism g_{n,v^c}^a on the experiment-selection set for fold v with TMLE targeting of the initial $Q_{n,v^c}^{\{0,s\}}$ on separate or pooled estimation sets, as described in Algorithm 1 below. In contrast, as used in the bias estimates, $Q_{n,v^c}^{\{0,s\},*}$ indicates training and targeting of the outcome regression using experiment-selection set data for fold v. All bias and ATE estimates are obtained using TMLE, which is a doubly-robust plug-in estimator that targets initial model fits to optimize the bias-variance tradeoff for the target parameter [40, 41]. In the case of the ATE, TMLE is asymptotically unbiased if either the outcome regression or the treatment mechanism are estimated consistently and is asymptotically efficient if both are estimated consistently [41]. A detailed description of targeted maximum likelihood estimation of the bias term $\hat{\Psi}_s^{\#}(P_n)$ may be found in Appendix A.2. The empirical selectors, based on experiment-selection set data for each fold, are then

$$s_n^{\star}(v^c) = \underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_{s,n,v^c}}}}{n} + (\hat{\Psi}_s^{\#}(Q_{n,v^c}^{\{0,s\},*}, Q_{n,v^c}^{s,*}))^2$$

$$s_n^{\star\star}(v^c) = \underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_{s,n,v^c}}}}{n} + (\hat{\Psi}_s^{\#}(Q_{n,v^c}^{\{0,s\},*}, Q_{n,v^c}^{s,*}) + \hat{\Phi}_s(Q_{n,v^c}^{NCO,*}))^2$$

where $Q^s = E[Y|S \in \{0,s\}, S, A, W]$, $Q^{NCO} = E[NCO|S \in \{0,s\}, A, W]$, and

$$D^*_{\Psi_{s,n,v^c}}(O,\bar{V}) = \frac{I(S\in\{0,s\},\bar{V}\neq v)}{P_n(S\in\{0,s\},\bar{V}\neq v)}\left(\left(\frac{I(A=1)}{g^a_{n,v^c}(A=1|W,S\in\{0,s\})} - \frac{I(A=0)}{g^a_{n,v^c}(A=0|W,S\in\{0,s\})}\right)\right.$$
$$(Y - Q_{n,v^c}^{\{0,s\},*}(S \in \{0,s\}, A, W))$$
$$\left. + Q_{n,v^c}^{\{0,s\},*}(S \in \{0,s\}, 1, W) - Q_{n,v^c}^{\{0,s\},*}(S \in \{0,s\}, 0, W) - \hat{\Psi}_s(Q_{n,v^c}^{\{0,s\},*})\right)$$

Algorithm 1 describes the overall estimation process for the experiment-selector CV-TMLE.

Algorithm 1 CV-TMLE for Data-Adaptive Experiment Selection

1: To ensure **Assumption 1**, **trim** data so no W values are not represented in RCT.

2: **Divide** $O^n = (O_1, ..., O_n)$ into V folds stratified on S with experiment-selection set $O^n_{v^c} = \{O_i : i = 1, ..., n, \bar{V}_i \neq v\}$ and estimation set $O^n_v = \{O_i : i = 1, ..., n, \bar{V}_i = v\}$.

3: **For** $v \in \{1, ..., V\}$,

1. **For all** $I(S \in \{0, s\})$ **subsets of** $O^n_{v^c}$ **experiment-selection sets**

- **Estimate:** Q^s_{n,v^c}, $Q^{\{0,s\}}_{n,v^c}$, Q^{NCO}_{n,v^c}, g^s_{n,v^c} §, g^a_{n,v^c}, $\frac{\hat{\sigma}^2_{D^*_{\Psi_{s,n,v^c}}}}{n}$
- **Use TMLE to estimate** $\hat{\Psi}^{\#}_s(Q^{\{0,s\},*}_{n,v^c}, Q^{s,*}_{n,v^c})$, $\hat{\Phi}_s(Q^{NCO,*}_{n,v^c})$
- **Select** experiment based on $s^\star_n(v^c)$, or $s^{\star\star}_n(v^c)$

4: **For all** O^n_v **estimation sets**

- **For all** $S \in \{0, s\}$
  1. **Pool** all $S \in \{0, s\}$ subsets of O^n_v across all v
  2. **Estimate** coefficient for TMLE update ϵ_s using logistic regression of binary or scaled-continuous Y on

$$H^*_s(A, W)^\dagger = \frac{I(A = 1)}{g^a_{n,v^c}(A = 1|S \in \{0, s\}, W)} - \frac{I(A = 0)}{g^a_{n,v^c}(A = 0|S \in \{0, s\}, W)}$$

  with offset $logit(Q^{\{0,s\}}_{n,v^c}(S \in \{0, s\}, A, W))$ pooled across all v based on initial regressions trained in experiment-selection sets.

- **For all** $v \in \{1, ..., V\}$
  1. **Select** the $I(S \in \{0, s^\star_n(v^c)\})$ subset of O^n_v
  2. Use $\epsilon_{s^\star_n(v^c)}$ to obtain targeted estimates¶

$$Q^{\{0,s^\star_n(v^c)\},*}_{n,v}(S \in \{0, s^\star_n(v^c)\}, 1, W) =$$
$$logit^{-1}(logit(Q^{\{0,s^\star_n(v^c)\}}_{n,v^c}(S \in \{0, s^\star_n(v^c)\}, 1, W)) + \frac{\epsilon_{s^\star_n(v^c)}}{g^a_{n,v^c}(A=1|S\in\{0,s\},W)})$$
$$Q^{\{0,s^\star_n(v^c)\},*}_{n,v}(S \in \{0, s^\star_n(v^c)\}, 0, W) =$$
$$logit^{-1}(logit(Q^{\{0,s^\star_n(v^c)\}}_{n,v^c}(S \in \{0, s^\star_n(v^c)\}, 0, W)) - \frac{\epsilon_{s^\star_n(v^c)}}{g^a_{n,v^c}(A=0|S\in\{0,s\},W)})$$

$$\hat{\Psi}_{s^*_n(v^c)}(Q^{\{0,s^\star_n(v^c)\},*}_{n,v}) =$$
$$\frac{1}{n}\sum_{i=1}^n \frac{I(S_i\in\{0,s^\star_n(v^c)\},\bar{V}_i=v)}{P_n(S\in\{0,s^\star_n(v^c)\},\bar{V}=v)}[Q^{\{0,s^*_n(v^c)\},*}_{n,v}(S_i \in \{0, s^\star_n(v^c)\}, 1, W_i) -$$
$$Q^{\{0,s^\star_n(v^c)\},*}_{n,v}(S_i \in \{0, s^\star_n(v^c)\}, 0, W_i)]$$

5: **Calculate** $\psi_n = \frac{1}{V}\sum_{v=1}^V \hat{\Psi}_{s^\star_n(v^c)}(Q^{\{0,s^\star_n(v^c)\},*}_{n,v})$

§: $g^s = P(S = 0|S \in \{0, s\}, A = 0, W)$

†: An alternative method that may be more stable in the context of practical positivity violations is to "target the weights" [42, 43, 44] by using clever covariate $H^*_s(A, W) = I(A = 1) - I(A = 0)$ and weights $\frac{I(A=1)}{g^a_{n,v^c}(A=1|S\in\{0,s\},W)} + \frac{I(A=0)}{g^a_{n,v^c}(A=0|S\in\{0,s\},W)}$.

¶: Re-scale $Q^{\{0,s^\star_n(v^c)\},*}_{n,v}$ to the original outcome scale if using scaled-continuous Y

1.5.1 Asymptotic distribution of the experiment-selector CV-TMLE

Next, we examine the asymptotic distribution of the ES-CVTMLE. Unlike the CV-TMLE for data-adaptive target parameter estimation developed by Zheng and van der Laan[14], the limit distribution of the ES-CVTMLE depends on the amount of bias introduced by a given real-world dataset. The finite sample challenge for selecting an optimal experiment depends on the magnitude of this true bias relative to the standard error of the ATE estimator, which in turn depends on the sample size. As noted by Yang et al.[10] for their elastic integrative analysis estimator, in order to understand the behavior of a selector in the context of this finite sample estimation challenge, we must understand the behavior of the selector when the bias is not fixed but rather dependent on the sample size. To accomplish this goal, define $P_{0,n}$ as the true data distribution dependent on n. In order to define the limit distribution, let us also define the following quantities.

Definitions Relevant for Asymptotic Distribution

$$Z_n(s,v) = \sqrt{n}(\hat{\Psi}_s(Q_{n,v}^{\{0,s\},*}) - \Psi_s(P_0)) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^n D^*_{\Psi_s,v}(O_i, V_i)$$
$$Z_n = (Z_n(s,v) : s = 1, ..., K, v = 1, ..., V) \sim Z = (Z(s,v) : s = 1, ..., K, v = 1, ..., V)$$

$$D^*_{\Psi_s,v}(O,\bar{V}) = \frac{I(S\in\{0,s\},\bar{V}=v)}{P(S\in\{0,s\},\bar{V}=v)}((\frac{I(A=1)}{g^a(A=1|W,S\in\{0,s\})} - \frac{I(A=0)}{g^a(A=0|W,S\in\{0,s\})})$$
$$(Y - Q^{\{0,s\}}(S \in \{0,s\}, A, W))$$
$$+Q^{\{0,s\}}(S \in \{0,s\}, 1, W) - Q^{\{0,s\}}(S \in \{0,s\}, 0, W) - \Psi_s(P_0))$$

$$Z_n^{\#}(s,v^c) = \sqrt{n}(\hat{\Psi}_s^{\#}(Q_{n,v^c}^{\{0,s\},*}, Q_{n,v^c}^{s,*}) - \Psi_s^{\#}(P_{0,n})) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^n D^*_{\Psi_s^{\#},v}(O_i, V_i)$$
$$Z_n^{\#} = (Z_n^{\#}(s,v) : s, v) \sim Z^{\#} = (Z^{\#}(s,v) : s, v)$$

$$Z_n^{\#+\Phi}(s,v^c) = \sqrt{n}(\hat{\Psi}_s^{\#}(Q_{n,v^c}^{\{0,s\},*}, Q_{n,v^c}^{s,*}) + \hat{\Phi}_s(Q_{n,v^c}^{NCO,*}) - (\Psi_s^{\#}(P_{0,n}) + \Phi_s(P_{0,n}))) \approx$$
$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (D^*_{\Psi_s^{\#},v}(O_i, V_i) + D^*_{\Phi_s,v}(O_i, V_i))$$
$$Z_n^{\#+\Phi} = (Z_n^{\#+\Phi}(s,v) : s, v) \sim Z^{\#+\Phi} = (Z^{\#+\Phi}(s,v) : s, v)$$

$$D^*_{\Psi_s^{\#},v^c}(O,\bar{V}) = \frac{I(S\in\{0,s\},\bar{V}\neq v)}{P(S\in\{0,s\},\bar{V}\neq v)}((\frac{I(S=0,A=0)}{g(S=0,A=0|S\in\{0,s\},W)})(Y - Q^s(S \in \{0,s\}, S, A, W))$$
$$-\frac{I(A=0)}{g^a(A=0|S\in\{0,s\},W)}(Y - Q^{\{0,s\}}(S \in \{0,s\}, A, W)) + Q^s(S \in \{0,s\}, S = 0, A =$$
$$0, W) - Q^{\{0,s\}}(S \in \{0,s\}, A = 0, W) - \Psi_s^{\#}(P_{0,n}))$$

$$D^*_{\Phi_s,v^c}(O,\bar{V}) = \frac{I(S\in\{0,s\},\bar{V}\neq v)}{P(S\in\{0,s\},\bar{V}\neq v)}((\frac{I(A=1)}{g^a(A=1|S\in\{0,s\},W)} - \frac{I(A=0)}{g^a(A=0|S\in\{0,s\},W)})$$
$$(NCO - Q^{NCO}(S \in \{0,s\}, A, W))$$
$$+Q^{NCO}(S \in \{0,s\}, 1, W) - Q^{NCO}(S \in \{0,s\}, 0, W) - \Phi_s(P_{0,n}))$$

$$D^*_{(\#+\Phi)_s,v^c}(O,\bar{V}) = D^*_{\Psi_s^{\#},v^c}(O,\bar{V}) + D^*_{\Phi_s,v^c}(O,\bar{V})$$

Next we consider the distribution of the standardized selectors, which are random variables that depend on the distribution of $Z_n^{\#}(s,v^c)$ or $Z_n^{\#+\Phi}(s,v^c)$. Multiplying the selector by n and adding and subtracting the true value of the bias yields a standardized selector

$$s_n^\star(v^c) = \underset{s}{argmin}\ \hat{\sigma}^2_{D^*_{\Psi_{s,n,v^c}}} + (Z_n^\#(s,v^c) + \sqrt{n}(\Psi_s^\#(P_{0,n})))^2$$
$$s_n^{\star\star}(v^c) = \underset{s}{argmin}\ \hat{\sigma}^2_{D^*_{\Psi_{s,n,v^c}}} + (Z_n^{\#+\Phi}(s,v^c) + \sqrt{n}(\Psi_s^\#(P_{0,n}) + \Phi_s(P_{0,n})))^2$$

Let $s_n^\star = (s_n^\star(v^c) : v)$ and $s_n^{\star\star} = (s_n^{\star\star}(v^c) : v)$ represent the multivariate standardized selectors applied across all experiment-selection sets. Let $P^*_{n,v}$ denote training of initial estimators of the outcome regression $Q^{\{0,s\}}_{n,v^c}$ and treatment mechanism g^a_{n,v^c} on experiment selection sets with TMLE targeting on separate or pooled estimation sets to generate $Q^{\{0,s\},*}_{n,v}$ for fold v. Let P^*_{n,v^c} denote training and TMLE targeting of the relevant outcome regressions for each bias parameter on experiment selection sets for fold v.

Theorem 1: *Under conditions of convergence of second-order remainders, consistency of EIC estimation, and a Donsker class condition for bias term estimation specified in Appendix A.3*, $s_n^\star(v^c)$ and $s_n^{\star\star}(v^c)$ approximate the limit processes $\bar{s}^\star(v^c)$ and $\bar{s}^{\star\star}(v^c)$ such that

$$\bar{S}^\star(v^c) \sim \underset{s}{argmin}\ \sigma^2_{D^*_{\Psi_{s,v^c}}} + (Z^\#(s,v^c) + \sqrt{n}\Psi_s^\#(P_{0,n}))^2$$
$$\bar{S}^{\star\star}(v^c) \sim \underset{s}{argmin}\ \sigma^2_{D^*_{\Psi_{s,v^c}}} + (Z^{\#+\Phi}(s,v^c) + \sqrt{n}(\Psi_s^\#(P_{0,n}) + \Phi_s(P_{0,n})))^2$$

and the standardized experiment-selector CV-TMLE,

$$\sqrt{n}(\psi_n - \psi_{n,0}) = H(Z^\#, Z, \Psi^\#(P_{0,n}))) + o_P(1)$$
$$\text{or } \sqrt{n}(\psi_n - \psi_{n,0}) = H(Z^{\#+\Phi}, Z, \Psi^\#(P_{0,n}), \Phi(P_{0,n})) + o_P(1)$$

converges to a mixture of normal distributions defined by the sampling process depicted in Definition 1 below. The Proof of Theorem 1 may be found in Appendix A.3.

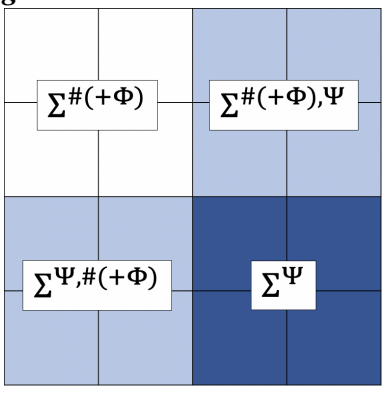
Definition 1. *Limit Distribution for Experiment-Selector CV-TMLE*

Across all $s = 0, ..., K$ and $v = 1, ..., V$, define the stacked vector of standardized experiment-selection set bias estimators and estimation set ATE estimators as

$$\tilde{Z} = (Z^\#, Z) \sim N(\vec{0}, \tilde{\Sigma}) \text{ or } \tilde{Z} = (Z^{\#+\Phi}, Z) \sim N(\vec{0}, \tilde{\Sigma})$$
where $\tilde{\Sigma}$ is depicted in Figure 1.1, and

$$\Sigma^\#((s_1, v_1^c), (s_2, v_2^c)) = E[D^*_{\Psi^\#_{s_1}, v_1^c}(O, \bar{V}) * D^*_{\Psi^\#_{s_2}, v_2^c}(O, \bar{V})]$$
$$\Sigma^{\#+\Phi}((s_1, v_1^c), (s_2, v_2^c)) = E[D^*_{(\#+\Phi)_{s_1}, v_1^c}(O, \bar{V}) * D^*_{(\#+\Phi)_{s_2}, v_2^c}(O, \bar{V})]$$
$$\Sigma^\Psi((s_1, v_1), (s_2, v_2)) = E[D^*_{\Psi_{s_1}, v_1}(O, \bar{V}) * D^*_{\Psi_{s_2}, v_2}(O, \bar{V})]$$
$$\Sigma^{\Psi,\#}((s_1, v_1), (s_2, v_2^c)) = E[D^*_{\Psi_{s_1}, v_1}(O, \bar{V}) * D^*_{\Psi^\#_{s_2}, v_2^c}(O, \bar{V})]$$
$$\Sigma^{\Psi,\#}((s_1, v_1), (s_1, v_1^c)) = 0$$

Figure 1.1: Covariance Matrix $\tilde{\Sigma}$



The limit distribution of the experiment-selector CV-TMLE is then defined by sampling from $\tilde{Z}$, calculating $\bar{s}^\star$ or $\bar{s}^{\star\star}$, and finally calculating

$$H(Z^{\#}, Z, \Psi^{\#}(P_{0,n})) = \frac{1}{V}\sum_{v=1}^{V}(Z(\bar{s}^{\star}(v^c), v))$$
$$or\ H(Z^{\#+\Phi}, Z, \Psi^{\#}(P_{0,n}), \Phi(P_{0,n})) = \frac{1}{V}\sum_{v=1}^{V}(Z(\bar{s}^{\star\star}(v^c), v)).$$

Asymptotic distribution of selector under varying magnitudes of bias

Table 1.1: Limit Distribution of Selector with Different Magnitudes of True Bias

Magnitude of Bias	Limit Distribution of Selector
Small $\sqrt{n}\Psi_s^{\#}(P_{0,n}) \xrightarrow{p} 0$	$\bar{S}^{\star}(v^c) \sim \underset{s}{argmin}\ \sigma^2_{D^*_{\Psi_s,v^c}} + (Z^{\#}(s, v^c))^2$
Intermediate $\sqrt{n}\Psi_s^{\#}(P_{0,n}) \xrightarrow{p} C$ where C is a constant	$\bar{S}^{\star}(v^c) \sim \underset{s}{argmin}\ \sigma^2_{D^*_{\Psi_s,v^c}} + (Z^{\#}(s, v^c) + C)^2$
Large $\sqrt{n}\Psi_s^{\#}(P_{0,n}) \xrightarrow{p} \infty$	$\bar{S}^{\star}(v^c) = 0$

As shown in Table 1.1, although the random selector depends on $\Psi^{\#}(P_{0,n})$, it converges to a limit distribution that does not depend on n, and which is known if bias is small, known up to a constant if bias is intermediate, and degenerate, selecting 0 with probability 1, if bias is large. To obtain inference for the experiment-selector CV-TMLE, we use Monte Carlo simulation to generate 1000 samples from the estimated limit distribution and define 95% confidence intervals based on the quantiles q^p of these samples as $\psi_n + (\frac{q^{0.025}}{\sqrt{n}}, \frac{q^{0.975}}{\sqrt{n}})$.

In the case where RCT-only is selected in all experiment-selection sets, we use influence curve-based variance estimates consistent with a standard CV-TMLE procedure, with confidence intervals estimated as $\psi_n \pm 1.96 * (\frac{1}{V}\sum_{v=1}^{V}\frac{\hat{\sigma}^2_{D^*s=0,n,v}}{n_{s=0}})^{1/2}$ [14, 15] where

$$D^*_{s=0,n,v} = (\frac{I(A=1)}{g^a_{n,v^c}(A=1|W,S=0)} - \frac{I(A=0)}{g^a_{n,v^c}(A=0|W,S=0)})(Y - Q^{s=0}_{n,v^c}(A,W)) + Q^{s=0}_{n,v^c}(1,W) - Q^{s=0}_{n,v^c}(0,W) - \hat{\Psi}_{s=0}(P^*_{n,v})$$

estimated among RCT estimation set observations for fold v. We use plug-in estimates for the relevant components of the efficient influence curves and for the bias terms in the selector. Because we overestimate bias for truly unbiased RWD, we expect the confidence intervals to be conservative in this case. Nonetheless, as shown through simulations below, this method of determining confidence intervals provides close to nominal coverage with both intermediate and large magnitudes of bias.

1.6 Simulations

The following simulation compares the ES-CVTMLE to an RCT-only t-test and an RCT-only CV-TMLE using the *tmle R* package [45], as well as four other data fusion methods described below across several magnitudes of external data bias and when the U-comparability assumption needed for bias adjustment with an NCO is false.

1.6.1 Data generation

We generate a small RCT (S=0) of 150 observations with probability of randomization to $A = 1$ of 0.67. The goal is to mimic a situation where, for ethical reasons, it is desirable to randomize more participants to active treatment. We also simulate three candidate real-world datasets $S \in \{1, 2, 3\}$ of 500 observations each, all with $A = 0$. Thus, no treatment is available outside the trial. Dataset $S = 1$ has the same data-generating distribution as the RCT except that all $A = 0$, so any apparent bias in $S = 1$ is due to finite sample variability. There are two unmeasured bias variables B_1 and B_2 that are deterministically 0 in $S = 0$ and $S = 1$ and are generated as follows in $S \in \{2, 3\}$. For this simulation, biased RWD could be included if it is approximately $\sqrt{\frac{\hat{\sigma}^2_{D^*_{\Psi_{s=0,n,v^c}}}}{n} - \frac{\hat{\sigma}^2_{D^*_{\Psi_{s\in\{0,2\},n,v^c}}}}{n}} = B = 0.21$. We then generate B_1 and B_2 as normally distributed random variables such that average total bias in $S = 2$ is $\approx B$ (intermediate bias) and in $S = 3$ is $\approx 5 * B$ (large bias). The outcome, Y is a function of both B_1 and B_2, while the NCO is only a function of B_1, so the U-comparability assumption is not true. Appendix A.4 contains further details regarding the data generating process and specifications for TMLE-based estimators used in this simulation.

1.6.2 Comparators

For each combination of $S = 0$ with one of $S \in \{1, 2, 3\}$, we compare our ES-CVTMLE with potential selectors $s_n^\star$ (**b2v**), $s_n^{\star\star}$ (**+nco**), and $s_n^{\star\star\star}$ (**nco only**) to four other data fusion estimators. These comparators were selected because they were developed for the context of augmenting a control arm of an RCT with external control data, they are commonly referenced, and they include methods for confidence interval construction. We introduce new versions of the test-then-pool approach originally described by Viele et al.[2] and the NCO-based difference-in-differences approach described by Sofer et al.[16] and Shi et al.[17], where our modifications use CV-TMLE estimators of the relevant parameters.

Test-then-pool

For the "test-then-pool" approach described by Viele et al.[2], a hypothesis test is conducted for a difference in the mean outcome of the trial controls and the mean outcome of the external controls. RCT and real-world data are combined if the null hypothesis is not rejected; if the null hypothesis is rejected, then the RCT data are analyzed without integration of RWD [2]. The original test-then-pool used an unadjusted estimator of the difference in mean outcome under treatment and control [2]. For the sake of comparison with other TMLE-based estimators, we include here both the method previously described, together with a minor extension that incorporates adjustment for baseline covariates for the sake of efficiency. For the unadjusted version, both the hypothesis test for including RWD and the treatment effect estimate are obtained using Welch's t-test with unequal variances. For the adjusted version, we first use CV-TMLE to estimate the ATE of S on Y among those with A=0 and decide to pool RCT and RWD if the 95% confidence interval for this estimate includes zero. We then obtain an estimate of the ATE of A on Y in the pooled or RCT-only sample, again using CV-TMLE. While the "test-then-pool" approach has been criticized for inappropriately including biased data due to low power of the test [21], a byproduct of this limitation is that the estimator is able to achieve large power gains when unbiased external controls are available. It is thus an interesting comparator as a high-risk, high-reward strategy for data fusion.

Meta-Analytic-Predictive priors

For comparison to a method of Bayesian Dynamic Borrowing, we use the *RBesT R* package [46] based on Schmidli et al.[3]. As described by Schmidli et al.[3] but modified for consistency with the above notation, $\theta_s = \Psi^0_{s,n,BDB}(P)$ is the mean outcome of controls in experiment $S \in \{0, s\}$. The prior distribution of θ_s is assumed to be Normal(μ, τ^2). τ is an estimate of the between-study heterogeneity that determines how much external control information is borrowed. For a continuous outcome, Weber et al.[46] recommend a Half-Normal$(0, \frac{\sigma}{2})$ prior distribution for τ, where σ is the standard deviation of the outcome estimated from external studies. Because the choice of the prior distribution of τ can impact results, Schmidli et al.[3] recommend conducting sensitivity analyses with different parameterizations of this distribution.

A sampling distribution of θ_s is generated using a Markov Chain Monte Carlo algorithm and approximated with a mixture of conjugate prior distributions [3]. To protect against non-exchangeability between external and trial controls, *RBesT* also provides a function to add a unit information prior component to this mixture [3, 46]. The weight of that vague prior must be specified by the researchers based on their beliefs regarding how likely the available control groups are to be exchangeable [3], with a suggested weight of 0.2 [46]. The control target parameter is estimated as the mean of the posterior distribution $E(\theta_s|O^n)$. The posterior distribution of the treatment target parameter is estimated as a mixture of conjugate distributions based on a weakly informative unit-information prior [46].

Negative control outcome (difference-in-differences approach)

Because our methods incorporate information from a negative control outcome, we also compare simulation results to a simple bias adjustment approach that is also based on an NCO. Multiple authors have noted that under the following assumptions, adjustment for bias using an NCO can be accomplished using a difference-in-differences approach [16, 17]. The first assumption is U-comparability, which states that all of the unmeasured factors that affect the A-Y relationship are the same as the unmeasured factors that affect the A-NCO relationship [24]. The second is "additive equi-confounding", which states that the unmeasured confounding has the same effect (on the additive scale) on the primary outcome as on the NCO [16, 17]. Under these assumptions, an estimator for the average treatment effect of A on Y for a given the experiment with $S \in \{0, s\}$ may be defined as $\hat{\Psi}^{DID}_s(P_n) = \hat{\Psi}_s(P_n) - \hat{\Phi}_s(P_n)$ [16, 17]. For a consistent comparison with the rest of our methods, we use CV-TMLE to estimate both parameters. The efficient influence curve of Ψ^{DID}_s is then $D^*_{\Psi^{DID}_s} = D^*_{\Psi_s} - D^*_{\Phi_s}$.

1.7 Simulation results

Table 1.2 shows the bias, variance, mean of the estimated variance, mean squared error (MSE), 95% confidence interval coverage, and power to detect the causal ATE (using $\alpha = 0.05$) across 1000 iterations of this simulation. The standard CV-TMLE analyzed using the RCT data alone had nominal coverage of 0.95 and power of 0.64. The RCT-only CV-TMLE had higher power than any of the unadjusted estimators, with the RCT t-test having coverage of 0.96 and power of 0.24.

The test-then-pool approaches were able to increase power as high as 0.93 for the TMLE-based test-then-pool when unbiased RWD were available. However, as bias in the RWD increased, the coverage suffered, dropping as low as 0.79 for the TMLE-based test-then-pool with $S = 2$ and 0.76 for the t-test based test-then-pool with $S = 3$. Test-then-pool is thus a high-risk, high-reward

approach to integrating observational and RCT data.

Because the U-comparability assumption is not true, the two methods that rely only on a bias estimate of the ATE of A on the NCO also exhibited decreased coverage. Bias in the NCO-based difference-in-differences approach increased as the bias in the available RWD increased, leading to coverage of 0.84 for the most biased RWD dataset (S=3). When we only considered the estimated ATE of A on NCO in the experiment-selector CV-TMLE ($s_n^{\star\star\star}$ (**nco only**)), coverage dropped as low as 0.87, which was lower coverage than when we also included $\hat{\Psi}^{\#}$ as an estimate of bias in the selector (discussed below).

With default specifications, *RBesT* [46] maintained coverage 0.94-0.97. Yet because this method does not adjust for covariates, power remained similar to the t-test, with higher power achieved by considering RWD with intermediate bias (power 0.32) than by considering unbiased RWD (power 0.29). Thus, while *RBesT* resulted in close to nominal coverage, this method had lower power than alternative estimators, including an adjusted CV-TMLE using only the RCT data. MSE was higher for the *RBesT* estimator than for any of the ES-CVTMLE estimators across all tested magnitudes of external bias.

Next, we examine the experiment-selector CV-TMLEs with the **b2v** and **+nco** selectors. With $S = 3$, these ES-CVTMLEs with selector $s_n^{\star}$ (**b2v**) or $s_n^{\star\star}$ (**+nco**) were approximately equivalent to the RCT-only CV-TMLE from the *tmle R* package. This makes sense because data with bias this large was rejected, in which case the ES-CVTMLE algorithm is equivalent to a traditional CV-TMLE from the RCT only. When unbiased external controls were available, coverage was 0.96 for $s_n^{\star}$ (**b2v**) and $s_n^{\star\star}$ (**+nco**), suggesting somewhat conservative confidence intervals consistent with the fact that estimated bias is included in the limit distribution sampling procedure despite truly being zero. Power increased compared to the RCT-only CV-TMLE in either case but was lower with $s_n^{\star}$ (**b2v**) at 0.74, compared to 0.83 for $s_n^{\star\star}$ (**+nco**), demonstrating the utility of including information from the estimated ATE of A on the NCO in the selector for incorporating truly unbiased external controls. With $S = 2$ (intermediate bias), coverage was 0.95 for $s_n^{\star}$ (**b2v**) and 0.92 for $s_n^{\star\star}$ (**+nco**), demonstrating that the experiment-selector CV-TMLE is able to maintain coverage close to 0.95 even with this challenging amount of bias and an imperfect NCO.

In this simulation, the ES-CVTMLE MSE was lower with either of the **b2v** or **+nco** selectors compared to the RCT-only CV-TMLE when considering $S = 1$ and lower or the same when considering $S = 2$ or $S = 3$. Of all the compared estimators, the ES-CVTMLE with the $s_n^{\star}$ (**b2v**) selector provided the largest power gains with unbiased RWD while maintaining 95% coverage across all tested magnitudes of bias. However, the ES-CVTMLE with the $s_n^{\star\star}$ (**+nco**) selector is the estimator that decreased MSE the most when unbiased RWD were available without increasing MSE when considering RWD with intermediate or large bias. If we were running this simulation to choose an estimator for a proposed trial in a context when excessive randomization to control is considered unethical but we still desire greater protection against biased conclusions than a purely observational analysis could achieve, we might choose the ES-CVTMLE with the $s_n^{\star\star}$ (**+nco**) selector because it was able to boost power substantially when appropriate external controls were available while keeping coverage close to nominal across a range of possible magnitudes of external bias, even when we did not have a perfect NCO.

Table 1.2: Results of Simulation - 1000 Iterations

Estimator (RWD)	Bias	Variance	Mean Est. Var.	MSE	Coverage	Power
RCT T-Test	0.005	0.206	0.219	0.206	0.96	0.24
RCT CV-TMLE	0.004	0.065	0.070	0.065	0.95	0.64
ES-CVTMLE $s_n^\star$ (**b2v**) (S=1)	0.003	0.054	0.058	0.054	0.96	0.74
ES-CVTMLE $s_n^\star$ (**b2v**) (S=2)	-0.026	0.065	0.061	0.065	0.95	0.71
ES-CVTMLE $s_n^\star$ (**b2v**) (S=3)	0.005	0.065	0.071	0.065	0.95	0.64
ES-CVTMLE $s_n^{\star\star}$ (**+nco**) (S=1)	0.005	0.045	0.044	0.045	0.96	0.83
ES-CVTMLE $s_n^{\star\star}$ (**+nco**) (S=2)	-0.028	0.059	0.052	0.060	0.92	0.76
ES-CVTMLE $s_n^{\star\star}$ (**+nco**) (S=3)	0.005	0.065	0.071	0.065	0.95	0.64
ES-CVTMLE $s_n^{\star\star\star}$ (**nco only**) (S=1)	0.004	0.028	0.034	0.028	0.97	0.92
ES-CVTMLE $s_n^{\star\star\star}$ (**nco only**) (S=2)	-0.152	0.036	0.038	0.059	0.87	0.95
ES-CVTMLE $s_n^{\star\star\star}$ (**nco only**) (S=3)	-0.037	0.089	0.068	0.090	0.91	0.67
TTP (CV-TMLE) (S=1)	0.004	0.037	0.029	0.037	0.93	0.93
TTP (CV-TMLE) (S=2)	-0.113	0.059	0.033	0.072	0.79	0.87
TTP (CV-TMLE) (S=3)	0.004	0.065	0.070	0.065	0.95	0.64
Diff-in-Diff (NCO) (S=1)	0.008	0.052	0.054	0.052	0.95	0.73
Diff-in-Diff (NCO) (S=2)	-0.040	0.054	0.054	0.056	0.94	0.79
Diff-in-Diff (NCO) (S=3)	-0.227	0.054	0.054	0.105	0.84	0.94
TTP (T-Test) (S=1)	-0.001	0.122	0.090	0.122	0.93	0.53
TTP (T-Test) (S=2)	-0.132	0.147	0.095	0.164	0.88	0.70
TTP (T-Test) (S=3)	-0.128	0.359	0.182	0.376	0.76	0.35
RBesT [46] (S=1)	-0.005	0.152	0.183	0.152	0.97	0.29
RBesT [46] (S=2)	-0.052	0.157	0.185	0.159	0.96	0.32
RBesT [46] (S=3)	-0.116	0.213	0.222	0.227	0.94	0.31

Caption: Mean. Est. Var.: Mean of variance estimates. S=1: unbiased RWD. S=2: RWD with intermediate bias. S=3: RWD with large bias. Power: Probability that confidence interval < 0 across 1000 iterations. TTP: Test-then-Pool. $s_n^\star = \underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + (\hat{\Psi}_s^\#(P_n))^2$, $s_n^{\star\star} = \underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + (\hat{\Psi}_s^\#(P_n) + \hat{\Phi}_s(P_n))^2$, $s_n^{\star\star\star} = \underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + (\hat{\Phi}_s(P_n))^2$

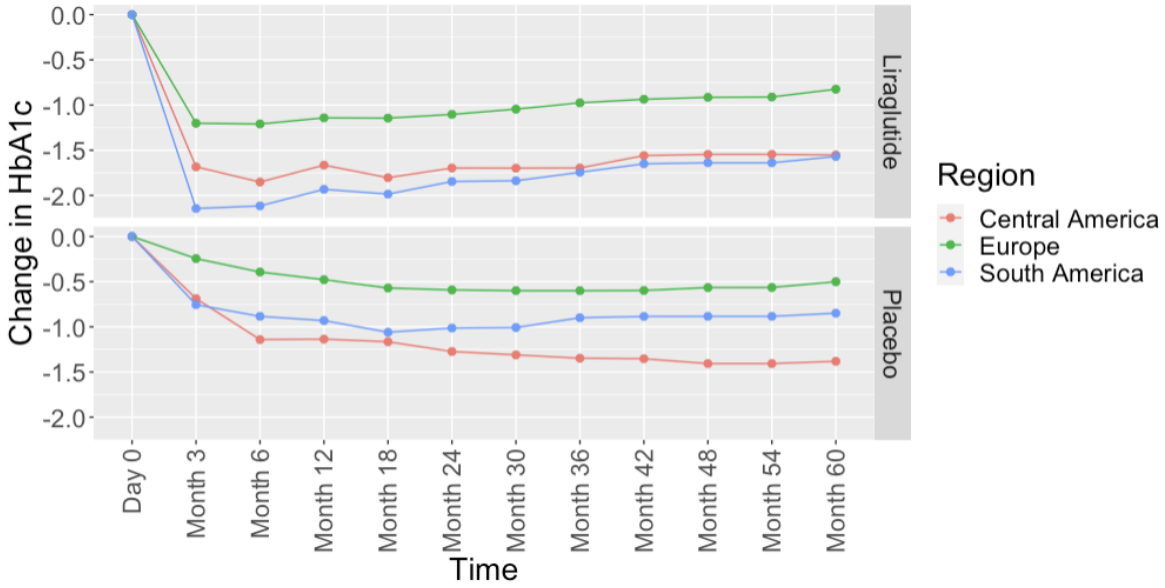
1.8 Real data application

Ultimately, the goal of the experiment-selector CV-TMLE is to facilitate integration of RCT and real-world data in order to boost RCT power without introducing bias. As an initial test case for this method, we have chosen an example where we have a fairly precise estimate of the true causal effect of interest from a well-powered multisite and multi-region RCT. We use these data to create a hypothetical scenario in which RCT data are only available from a subset of participants from one region —- resulting in an under-powered trial — but candidate control arm-only data are available from other regions (mimicking RWD of varying quality). We use this scenario to evaluate the ability of our proposed methods and others to recover the initial RCT effect estimate.

To create such a scenario, we use de-identified data from the LEADER trial (Clinical Trial NCT01179048). Initially reported by Marso et al.[47], this study evaluated the effect of an injectable (subcutaneous) glucagon-like peptide-1 receptor agonist, liraglutide, on a primary combined outcome of cardiovascular death, nonfatal myocardial infarction, or nonfatal stroke. The sample size for LEADER was 9340 patients. Because this trial was designed to evaluate relatively long-term and rare outcomes, the sample size was large enough to estimate the effect of liraglutide versus placebo (both added to standard of care therapy with oral antihyperglycemic drugs (OADs) and/or insulin) on glycemic control (measured by hemoglobin A1c (HbA_{1c})) with great precision.

LEADER encouraged trial clinicians to optimize standard of care diabetes regimens beyond the addition of liraglutide or placebo in order to achieve a target HbA_{1c} of $\leq 7\%$ for all trial participants [48]. We thus would expect not only to see a difference in change in HbA_{1c} between the liraglutide and placebo arms but also to see a change in HbA_{1c} from baseline in the placebo arm due to modifications in patients' baseline diabetes regimens. We would expect the major driver of change in HbA_{1c} in the placebo arm to be HbA_{1c} at baseline.

Figure 1.2: Change in HbA_{1c} by Trial Arm and Region Over Time

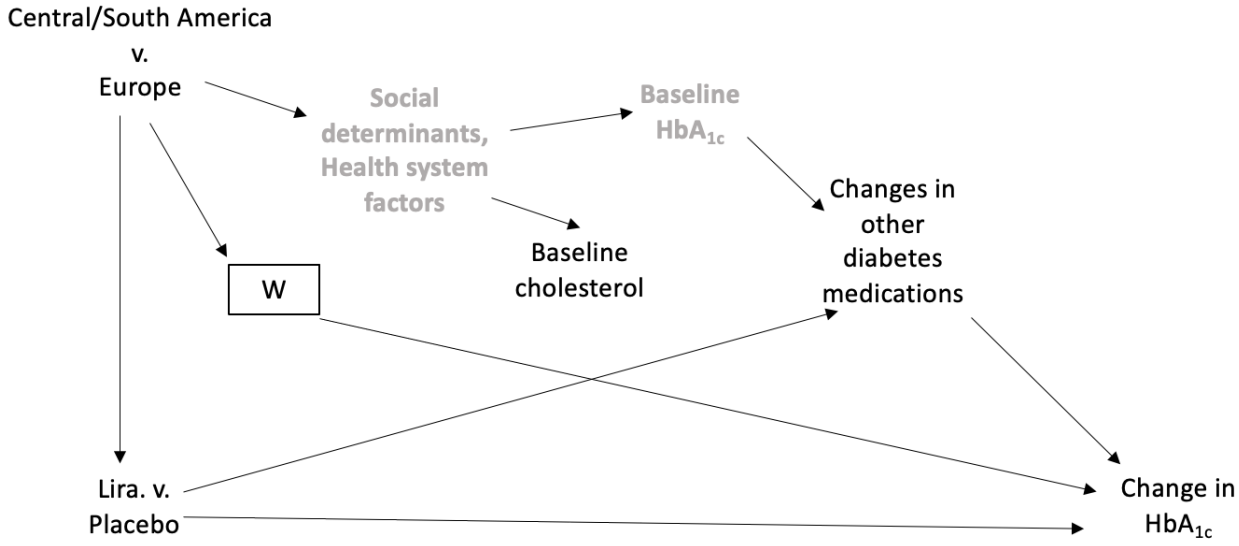


As shown in Figure 1.2, change in HbA_{1c} differed by study region, with the largest average changes in both the liraglutide and placebo arms taking place in the Central and South American groups. Average baseline HbA_{1c} was also higher in Central/South America (9.29) compared to

Europe (8.31). If we were to mimic a small RCT by taking a limited sample of patients from Central and South America and then augment the control arm with external controls from Central and South America, we would expect those individuals who were randomized to placebo from within the same region to be unbiased controls. However, if we were to augment the small Central/South America RCT with external controls from Europe, and if we treated baseline HbA_{1c} as an unmeasured factor that causes the differences in placebo group outcomes by region, we would expect the following. The treatment arm would only contain subjects from Central and South America, with a relatively large average decrease in HbA_{1c}. The addition of European controls to the placebo arm would lead to a smaller average change in HbA_{1c} among all controls, leading to an overestimate of the effect of liraglutide compared to placebo on glycemic control compared to the effect estimate from the full Central/South America LEADER subset.

This set-up implies the following directed acyclic graph:

Figure 1.3: Causal Graph for Analysis of LEADER Data



Lira.: Liraglutide

Based on our data set-up, region (Central/South America or Europe) affects treatment because members of the Central/South America group may receive liraglutide or placebo, and participants from Europe may only receive placebo. As we have noted, region also affects change in HbA_{1c} in the placebo arm. Because average baseline HbA_{1c} was higher and average improvement in HbA_{1c} was larger for the Central/South America compared to European subgroups, this suggests that on average, baseline diabetes regimens may have been less adequate in the Central/South America LEADER sample. In reviews of barriers and facilitators for diabetes management in Latin America, Blasco-Blasco et al.[49] and Aviles-Santa et al.[50] cite access to healthcare, limitations in health system resources, and social determinants of health as challenges that impede optimal glycemic control for many people. While these factors vary by country, differences in such underlying barriers between the Central/South American and European subgroups of LEADER could explain at least part of the noted difference in average baseline HbA_{1c}.

We also have access to the following baseline covariates, W: age, sex, smoking status (never, former, or current), diabetes duration, whether the patient is insulin naive at baseline, eGFR, and

BMI. Based on this DAG, we would expect baseline HbA_{1c} and W to block all paths from region to the outcome, other than the path through treatment, but we will treat HbA_{1c} as unmeasured.

The last ingredient for our analysis is an appropriate negative control. As shown in our causal graph, we may hypothesize that regional differences in health care for patients with metabolic syndrome causing inadequate control of HbA_{1c} may also lead to inadequate control of cholesterol. This hypothesis is supported by Venkitachalam et al.[51]'s finding that both country-level health systems factors and economic development metrics were significantly associated with prevalence of elevated cholesterol among patients with a history of hyperlipidemia from thirty-six countries. If this hypothesis is true, baseline cholesterol may serve as a negative control variable given that it would be associated with unmeasured factors hypothesized to cause differences in the placebo arm change in HbA_{1c} by region while not being affected by liraglutide administered post-baseline. Note also that we expect improvements in the adequacy of the baseline medication regimen to lead to smaller improvements in HbA_{1c} during the trial and also to be associated with lower levels of baseline cholesterol. By defining our outcome as improvement in HbA_{1c}, we satisfy our goal of defining a negative control variable that should be affected by the unmeasured bias in the same direction as the true outcome.

Our observed data thus consist of $O = (S, W, C, A, Y)$, where the W covariates are defined above, C is baseline cholesterol level in mmol/L, A is a binary indicator of liraglutide versus placebo, and Y is improvement in HbA_{1c} from baseline to study month 12. S is an indicator of study: 1 for the Central/South American "RCT" sample (random sample of 150 participants), 2 for extra controls from Central/South America (random sample of 500 participants not included in study 1), and 3 for extra controls from Europe (random sample of 500 participants). For clarity, we will refer to study 1 as C/S, a combination of studies 1 and 2 as C/S+, and a combination of studies 1 and 3 as Eu+. In order to demonstrate the case where we would like to increase the number of patients receiving the intervention of interest in our "RCT", we select $S = 1$ participants with a probability of 0.67 of having been in the liraglutide arm and 0.33 of having been in the placebo arm.

Overall missingness for change in HbA_{1c} was 6%. Missingness for baseline cholesterol, which was treated as an outcome for the estimate of the ATE of A on negative control in the selector but was treated as a baseline variable in the TMLE for the ATE of A on Y, was 2%. Outcome missingness was handled with inverse probability weights, consistent with the *tmle* package [45]. Specifically, we define a binary variable Δ that indicates an outcome was not missing. Clever covariates for all TMLEs were then modified to include the missingness indicator in the numerator and missingness mechanism in the denominator. For example, the clever covariate for the ATE was modified as $H(A, W) = \frac{\Delta(2*A-1)}{g(\Delta=1|A,W)g(A|W)}$. Missingness for baseline covariates, which was less than 0.1% for all W variables, was imputed using the *R* package *mice: Multivariate Imputation by Chained Equations* [52] separately for each study.

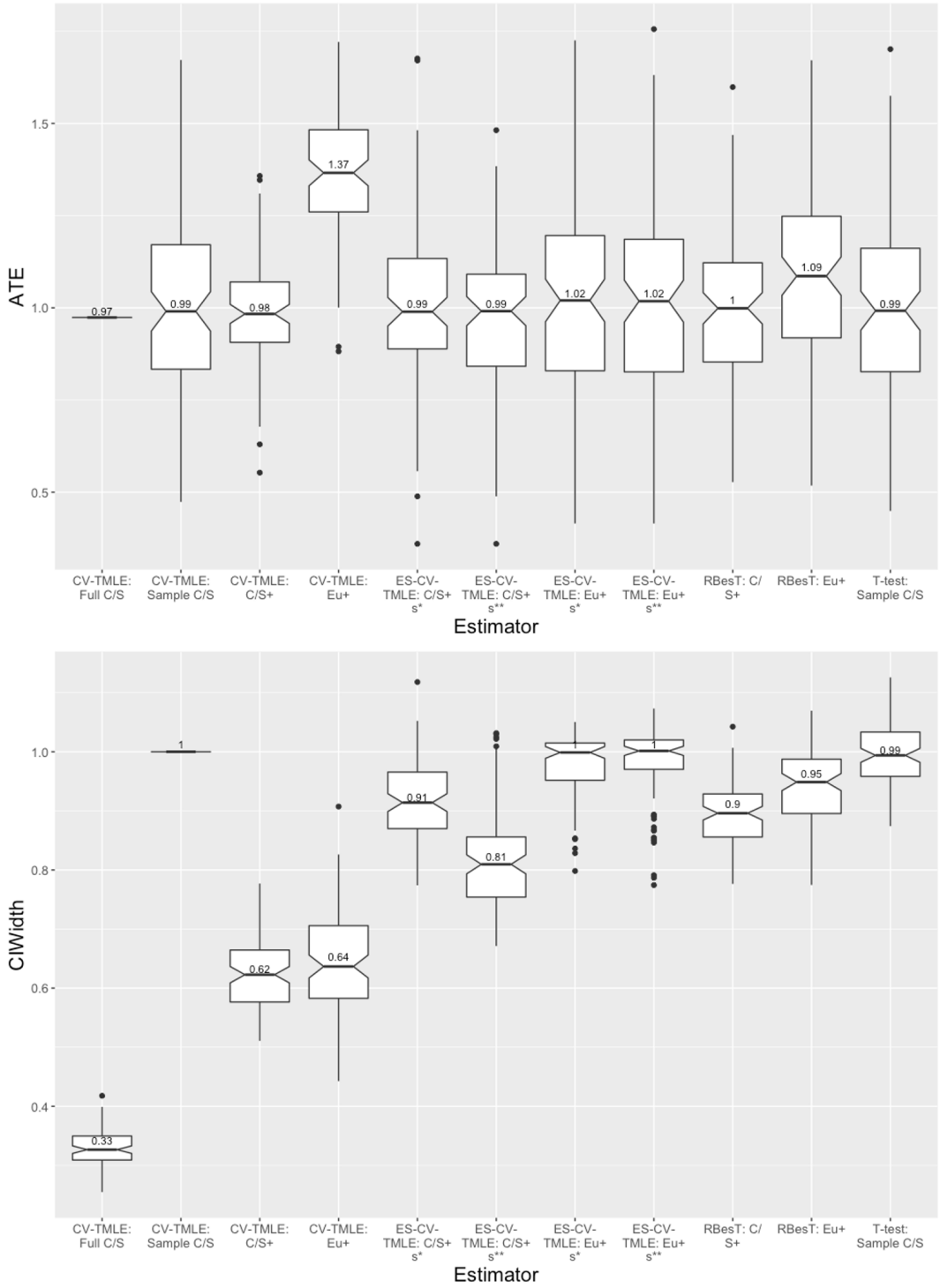
Our desired target causal parameter is the average treatment effect of liraglutide versus placebo on improvement in HbA_{1c} from baseline to 12 months in Central/South America. Due to randomization within the LEADER trial, this target parameter should be identifiable from dataset C/S and C/S+ but not from Eu+ without adjustment for baseline HbA_{1c}. We compare the following estimators: a CV-TMLE from the *tmle* package [45] using C/S only, the experiment-selector CV-TMLE considering C/S+ or considering Eu+, the *RBesT* package [46] considering C/S+ or considering Eu+, and a t-test using C/S only. To further demonstrate what could happen if the ATE were esti-

mated from data that includes biased controls without any evaluation of whether bias is present, we also include standard CV-TMLEs based on the C/S+ and Eu+ datasets. We run this analysis 100 times with different random seeds.

For the TMLEs, we use the following specifications. We employ a discrete Super Learner for all outcome regressions with a library consisting of linear regression [53], lasso regression (via *R* package *glmnet* [54]), and multivariate adaptive regression splines [55]. When considering only $S = 1$, we use the true randomization probability of 0.67 for $P(A = 1)$. When external controls are considered, we use a discrete Super Learner with library consisting of logistic regression and lasso regression for the treatment mechanism. Because missingness was low, for the missingness mechanism we use a linear model adjusting only for treatment unless the number of missing observations is less than five, in which case we employ an intercept only adjustment. We also use the *tmle* package defaults of fitting a CV-TMLE, using a logistic fluctuation, and targeting the weights, as described above.

1.8.1 Results of analysis of LEADER data

Figure 1.4: Estimated ATE of Liraglutide v. Placebo on Improvement in HbA_{1c} by Estimator



Caption: Boxplots of ATE and Relative confidence interval (CI) width with medians labeled. Relative width of CI compared to RCT CV-TMLE from sample C/S. Full C/S: Full Central/South America sample from LEADER trial (sample size 1182). C/S: Central/South American sample "RCT" (sample size 150). C/S+: C/S plus 500 additional controls from Central/South America. Eu+: C/S + 500 additional controls from Europe. ES-CV-TMLE: Experiment-selector CV-TMLE. CV-TMLE: Standard CV-TMLE from *tmle* package [45].

Figure 1.4 shows the ATE point estimates and relative confidence interval (CI) widths for each

estimator compared to the standard CV-TMLE from the small C/S "RCT" sample for 100 iterations of this real data analysis. The point estimate for the CV-TMLE from the full LEADER Central/South America subgroup was 0.97. Because, without adjustment for baseline HbA_{1c}, the remaining W covariates are not very predictive of change in HbA_{1c}, the point estimates and confidence interval widths for the C/S "RCT" sample CV-TMLE were both similar to those from a t-test from C/S. The narrowest CI was from the full LEADER Central/South America RCT CV-TMLE, followed by the standard CV-TMLEs run on C/S+ and Eu+.

Using a standard CV-TMLE for the C/S+ datasets, the median point estimate was similar to the full LEADER Central/South America estimate at 0.98. Yet without adjustment for baseline HbA_{1c}, the median ATE estimate for the standard CV-TMLE with Eu+ was severely biased at 1.37. This example demonstrates what could happen in this analysis if we did not know about the regional differences in baseline HbA_{1c} and decided to augment a Central/South America RCT with European controls with no analysis of bias.

Median (first and third quartile) values of the ATE estimate were 0.99 (0.83,1.17) for the C/S-only CV-TMLE. The point estimates from the *RBesT* package [46] were similar though less variable compared to the C/S subgroup estimates when dataset C/S+ was considered at 1.00 (0.85,1.12), and confidence intervals narrowed to a median of 0.9 compared to the C/S-only CV-TMLE. The experiment-selector CV-TMLEs considering C/S+ produced similar but slightly less variable results with median (first and third quartile) values of 0.99 (0.89,1.13) for $s_n^\star$ (**b2v**) and 0.99 (0.84,1.09) for $s_n^{\star\star}$ (**+nco**). The median relative confidence interval width for the $s_n^\star$ (**b2v**) selector was 0.91, while the median relative confidence interval width for the $s_n^{\star\star}$ (**+nco**) selector was 0.81. These results are consistent with the simulation results demonstrating that when unbiased external controls are added, the ES-CVTMLE with the $s_n^{\star\star}$ (**+nco**) selector leads to larger increases in power compared to the ES-CVTMLE with the $s_n^\star$ (**b2v**) selector or the *RBesT* package [46] with default settings. The relative confidence interval width for the experiment-selector CV-TMLE with C/S+ and selector $s_n^{\star\star}$ (**+nco**) is about half way between the CI width for the small C/S "RCT" sample and the CI width for the standard CV-TMLE from C/S+ with no assessment for bias, but this is the price paid for keeping the ATE estimates similar to estimates from the C/S "RCT" sample when biased Eu+ controls are considered.

When the Eu+ datasets were considered, the *RBesT* [46] median (first and third quartile) point estimates were shifted slightly upwards to 1.09 (0.92,1.25), with a median confidence interval width relative to the C/S-only CV-TMLE of 0.95. With the Eu+ datasets, the median (first and third quartile) of the ATE estimates from the experiment-selector CV-TMLE were 1.02 (0.83,1.20) with the $s_n^\star$ (**b2v**) selector and 1.02 (0.83,1.19) with the $s_n^{\star\star}$ (**+nco**) selector. The experiment-selector CV-TMLEs considering Eu+ had relative confidence interval widths that were similar to the C/S-only standard CV-TMLE at a median of 1 for either selector. These results suggest that the ES-CVTMLEs were less influenced by the biased external data than the *RBesT* [46] estimates were and demonstrate the relative robustness of the ATE estimates from the experiment-selector CV-TMLE when potentially biased external controls are considered.

1.9 Discussion

We introduce a novel cross-validated targeted maximum likelihood estimator that aims to select the experiment (RCT or RCT plus external controls) that optimizes the bias-variance tradeoff for the causal average treatment effect. To address the challenge that the selector may remain random

asymptotically with small to intermediate magnitudes of external bias, we develop an algorithm for confidence interval construction that samples from the estimated limit distribution and that includes an estimate of the bias in this sampling process. Through simulations, we demonstrate that we are able to improve power compared to a standard CV-TMLE from the RCT only when unbiased external controls are available and maintain coverage close to 95% with intermediate to large magnitudes of bias. In an analysis of the ATE of liraglutide versus placebo on improvement in 12 month HbA_{1c} from the LEADER trial, we also demonstrate the ability of the experiment-selector CV-TMLE to include external controls and narrow confidence intervals when additional unbiased controls are available and to reject biased external controls in the majority of iterations, maintaining similar confidence interval widths and point estimates compared to the sample "RCT"-only CV-TMLE.

The purpose of the experiment-selector CV-TMLE is to provide an estimator that is robust to varying magnitudes of bias from a combined RCT-RWD analysis when, as may frequently happen in partially observational studies, we are not certain whether the mean exchangeability assumption or U-comparability assumptions are true. Many existing methods rely explicitly on these assumptions. Others either rely on a comparison of mean outcomes or effect estimates for RCT versus external participants [2, 11, 13, 12] or evaluate the effect of treatment on an NCO [17], but not both. Because bias must be estimated from the data, attempts to optimize the bias-variance tradeoff may either inadvertently exclude truly unbiased external data or include external data with a magnitude of bias that may impact causal coverage. By including an estimate of the ATE of treatment on a negative control outcome in our selector, we are able to more frequently include truly unbiased external controls in our analysis. Yet we do not require the NCO to be perfect and show improved coverage compared to an NCO-based bias-adjustment approach when the U-comparability assumption does not actually hold. We thus aim to improve on existing methods by incorporating information from both an estimated causal gap and from a negative control outcome to maximize our ability to select an optimal experiment for analyzing a causal ATE. Another advantage of the experiment-selector CV-TMLE is that it attempts to learn how much external information to include only from the data, rather than requiring a researcher to specify a level of confidence in the external controls as is required in some Bayesian dynamic borrowing approaches [3] or to specify the value of a tuning parameter as is required by some frequentist approaches [11, 10].

The largest limitation of the experiment-selector CV-TMLE is that, because we cannot guarantee 95% coverage, the performance may depend on characteristics of the proposed analysis. Once again, this limitation is not unique to our estimator, as other data fusion estimators have demonstrated either increases in type 1 error or relative MSE or decreases in power with differing magnitudes of external data bias [5, 2, 11, 13, 18, 19, 20, 10, 12]. Yet because of this limitation, it would be important to conduct an outcome-blind simulation that is as true to a proposed study as possible, prior to implementing this estimator in a different context. Outcome-blind simulations can address differences in estimator performance for differing study characteristics and estimator specifications, such as different RCT and RWD sample sizes, the relative predictiveness of the covariates for the outcome, the candidate algorithms, the number of cross-validation folds, and the outcome type. For example, it is possible that for a given study design, the optimal bias-variance tradeoff across varying magnitudes of potential bias could actually be achieved by adding a smaller number of external controls than are available. A future version of the selector could consider adding different numbers of external controls based on, for example, increasing numbers of propensity-score matched external participants. In future work, we also intend to evaluate the experiment-selector CV-TMLE in a wider variety of contexts, including extending the methods to include time-to-event outcomes.

This real data analysis allowed the opportunity to test the experiment-selector CV-TMLE in a setting where we understand the "unmeasured" factors causing external controls to be biased or unbiased. In the future, we intend to test this method when attempting to combine real electronic health records data with a small RCT sample, again with the aim of replicating the full trial results. While this approach may prove viable in some settings, we suggest that in order to optimize the probability both that RWD is included and that bias is truly minimal, in many cases a preferable approach will be to prospectively specify a hybrid RCT-RWD study, allowing protocols and measurements to be made as similar as possible. We do not intend these methods to be a replacement for a traditional randomized controlled trial when it is feasible to run one for the sake of evaluating the efficacy of a new drug that has yet to be approved. Yet we hope that the experiment-selector CV-TMLE may ultimately be able to provide evidence to support conclusions from under-powered RCTs conducted for rare diseases, to allow randomization of more patients to the intervention arm for medications evaluated for severe diseases with few treatment options, and to contribute robust evidence to the evaluation of previously approved drugs for new populations and indications.

2 The EScvtmle R Package

2.1 Introduction

In recent years, studies that integrate randomized controlled trial (RCT) data with external data to estimate a causal effect have become increasingly popular. For example, a recent case study from the Food and Drug Administration (FDA)'s Complex Innovative Trial Design (CID) program proposed a standard RCT for estimation of the effect of an experimental medication on progression-free survival for patients with diffuse large B-cell lymphoma yet was not adequately powered for the outcome of overall survival [56]. To improve efficiency for estimating the effect on this secondary outcome, the Sponsor proposed augmenting the control arm of their study with data from the control arm of another clinical trial [56]. In another example, Brunner et al.[57] report a small RCT (sample size 93 children with systemic lupus erythematosus (SLE)) of the effect of Belimumab on the SLE Responder Index (SRI4) response rate. The authors note that this "study [is] not powered for statistical testing" [57], because childhood-onset SLE is so rare, but consider integrating information from adult trials of Belimumab to improve study power.

These two trials demonstrate common motivations for conducting such hybrid randomized-external data studies. For rare diseases, it may be challenging to recruit sufficient participants to estimate — with adequate precision — the effect of a new medication, particularly when the outcome is rare, as well [4]. In other cases, patients or the parents of pediatric patients may wish to minimize the probability that trial participants are randomized to control status because prior data from trials in adults or for a different form of the same drug suggest that the active treatment is very likely to be beneficial [2]. Conversely, for severe diseases with limited existing treatment options, participants may desire a higher probability of randomization to any treatment that has the possibility of improving on the existing standard of care [58]. In such cases, including external data can decrease the number of required total or control participants. Because the FDA's Real World Evidence Program considers the incorporation of non-randomized data in the regulatory approval process [59] and the FDA's CID Program encourages the use of innovative designs [56], we expect the number of such hybrid RCT-external data studies to increase over time.

In order to protect against biased conclusions when external data are observational or from a different population or setting, there has been growing interest in developing methods that estimate causal bias in order to decide if and how to integrate external with randomized data. Regardless of whether a Bayesian (e.g., [8, 3]) or Frequentist (e.g., [10, 11, 12, 13, 60]) approach is taken, these methods estimate bias or determine the degree of information borrowing based on the difference in mean outcomes, conditional mean outcomes, or parameter estimates between the RCT and external data sources. Yet finite sample variability in bias estimation leads to a tradeoff between the probability that unbiased data are integrated, leading to an improvement in efficiency, and the probability that biased external data are excluded, protecting against increases in type 1

error [13]. While no method can both guarantee nominal coverage regardless of the potential magnitude of external data bias and gain efficiency over an estimator using the RCT alone [11], we recently developed the experiment-selector cross-validated targeted maximum likelihood estimator (ES-CVTMLE) methodology to address some of the limitations of existing methods for analyzing hybrid randomized-external data studies [60].

First, to modify the balance of the tradeoff between efficiency gains with unbiased data and type 1 error control with biased data, some estimators require the user to specify the value of a tuning parameter (e.g., [3, 11]). Rather than relying on such a subjective process, the ES-CVTMLE incorporates a second objective assessment of bias by estimating the average treatment effect (ATE) on a negative control outcome (NCO). A second advantage is that the ES-CVTMLE uses targeted maximum likelihood estimation (TMLE) to estimate both the bias terms and the target parameter — in this case the average treatment effect (ATE). Because TMLE is an efficient, doubly-robust plug-in estimator, where initial estimates of the outcome regression and treatment mechanism are estimated non-parametrically using the SuperLearner machine learning prediction algorithm [61], TMLE minimizes statistical assumptions, optimizes the statistical bias-variance tradeoff for the target parameter, and respects the bounds of the statistical model [41]. Furthermore, CV-TMLE promotes accurate inference for data-adaptive target parameters by separating parameter definition and nuisance function estimation from effect estimation using cross-validation [62, 15].

Several existing software packages conduct TMLE but are not set up to handle the extra challenges of data-adaptive experiment selection. The *tmle R* package performs TMLE or CV-TMLE to estimate the effect (ATE and relative risk, among other parameters) of a binary treatment on an outcome [45]. The *ltmle R* package [63] implements TMLE for longitudinal data structures. The *tlverse* software ecosystem [64] extends the functionality of the original two packages to complex parameters including the effects of optimal individualized treatment regimes, stochastic treatment regimes, and mediation analysis. Nonetheless, none of these packages handle the two primary challenges of the ES-CVTMLE methodology: causal bias estimation and variance estimation for the ES-CVTMLE, which converges to a mixture of normal distributions [60].

Reliable software is necessary to ensure that complex statistical methods are made available for broad use in a manner that is reproducible and transparent. To increase accessibility of the experiment-selector CV-TMLE methodology, we have developed the *EScvtmle R* package, available on CRAN at `https://cran.r-project.org/web/packages/EScvtmle/index.html`. The purpose of this chapter is to describe the use of this package, including handling of different situations such as missing outcomes and cluster-randomized datasets. We also introduce a minor extension from the methods described in Chapter 1 to allow for active treatment in the external data. Then, we use a publicly available real data example (included with the package) from the WASH Benefits Bangladesh cluster randomized controlled trial [65] to demonstrate the use of the *EScvtmle R* package. Finally, we discuss plans for future extensions of the package.

2.2 Statistical background

Full details of the ES-CVTMLE methodology may be found in Chapter 1. We provide an abbreviated review of this method here. Throughout, we will use A to refer to the binary intervention of interest with $A = 1$ for active treatment and $A = 0$ for control, W to refer to baseline covariates, Y to refer to the primary outcome, and S to refer to the study in which an individual participated, with $S = 1$ indicating the primary RCT. We may then define multiple possible experiments that

could be run as $S = 1$ (RCT-only) or $S \in \{1, s\}$ as RCT combined with external dataset s. While the experiment-selector CV-TMLE may theoretically consider any number of experiments, the *ES-cvtmle* package is set up to compare two potential experiments, where participation in the external dataset is indicated by $S = 0$.

With counterfactual outcomes [28] an individual would have had if prescribed treatment $A = a$ denoted as Y^a, the full set of endogenous variables is $X^F = (S, W, A, Y^0, Y^1)$. Each variable has its own exogenous error such that $U = (U_S, U_W, U_A, U_Y)$. The distribution of the endogenous and exogenous variables is then denoted by $P_{U,X}$. The observed data are n i.i.d. draws $O = (S, W, A, Y)$ from the true observed data distribution P_0.

Our overall goal is to estimate the average treatment effect (ATE) of A on Y. If similar populations are represented in the RCT and external data, we may be happy with estimating the ATE for either of our two experiments with $S = 1$ or with $S \in \{0, 1\}$ — whichever could be estimated with a more optimal bias-variance tradeoff. The ATE of a generic experiment with $S \in \{1, s\}$ is given by

$$\Psi^F_s(P_{U,X}) = E_{W|S\in\{1,s\}}[E(Y^1 - Y^0|W, S \in \{1, s\})].$$

As for other data-adaptive target parameters, we wish to avoid using the same data to both define and estimate our parameter of interest [62, 15], so the ES-CVTMLE separates experiment-selection from effect estimation using cross-validation. To accomplish this task, as shown in Figure 2.1, the ES-CVTMLE divides the data into V folds. In each fold, $1/V$ of the data is used as the estimation set and assigned variable $\bar{V}_i = v$, and $(V - 1)/V$ of the data is used as the experiment-selection set and assigned $\bar{V}_i \neq v$. We may then define the empirical distribution of estimation set observations for a given fold as $P_{n,v}$ and the empirical distribution of experiment-selection set observations for a given fold as P_{n,v^c}.

Our overall target parameter is the average of the ATEs for the selected experiment — denoted $s^\star_n$ – in each experiment selection set, v^c:

$$\psi_0 = \frac{1}{V}\sum_{v=1}^{V} \Psi^F_{s^\star_n(v^c)}(P_{U,X}).$$

Under causal identification assumptions of positivity (see [30, 31]) and mean exchangeability and generalizability (see [32, 33]), this target parameter is equal to the statistical estimand

$$\psi_{n,0} = \frac{1}{V}\sum_{v=1}^{V} \Psi_{s^\star_n(v^c)}(P_0)$$

where by the g-computation formula [29]

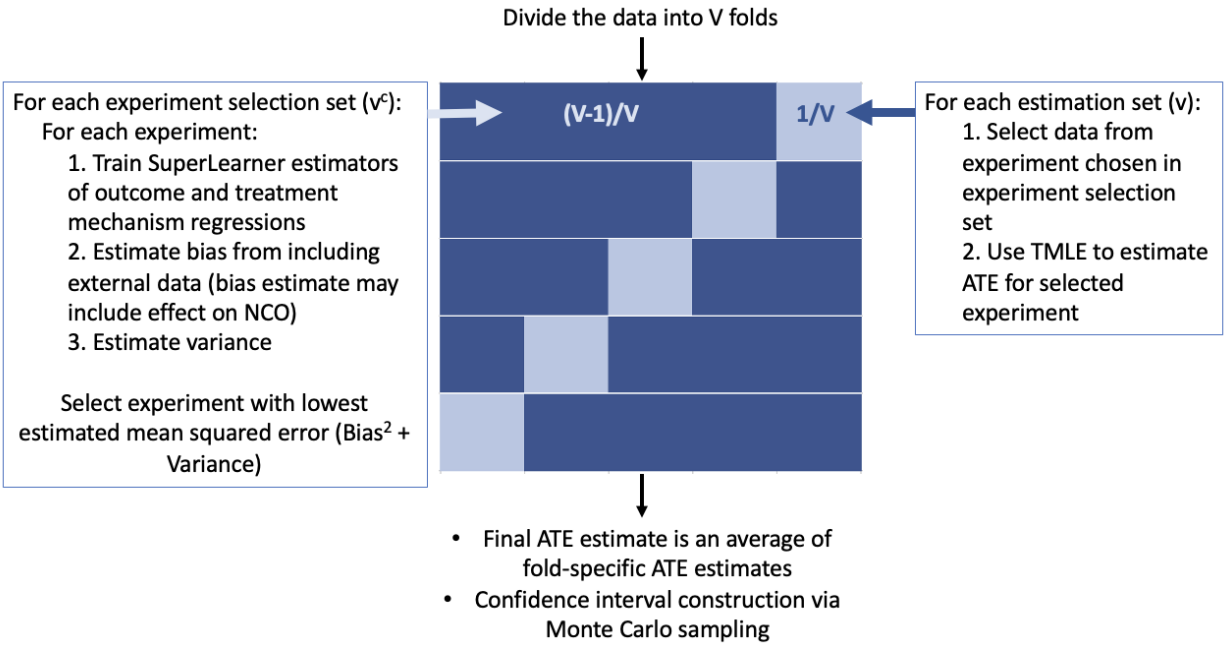
$$\Psi_s(P_0) = E_{W|S\in\{1,s\}}[E_0[Y|A = 1, W, S \in \{1, s\}] - E_0[Y|A = 0, W, S \in \{1, s\}]].$$

The ES-CVTMLE estimates $\psi_{n,0}$ using the following procedure.

2.2.1 Experiment-selection sets

As depicted in Figure 2.1, for each experiment-selection set, for each experiment, the *EScvtmle* package estimates the causal bias from inclusion of external data and the variance of the ATE estimator for that experiment and selects the experiment with the smallest estimated mean squared

Figure 2.1: Diagram depicting experiment-selector CV-TMLE procedure



error (squared bias plus variance). We review these steps in greater detail below.

2.1.1 Training of outcome and treatment mechanism regressions

In order to estimate the bias and variance of the ATE estimator for each experiment, we must first estimate $g_n(A|S \in \{1, s\}, W) = P_n(A = a|S \in \{1, s\}, W)$ and $Q_n(S \in \{1, s\}, A, W) = E_n[Y|S \in \{1, s\}, A, W]$. These outcome and treatment mechanism regressions will be used in three ways: 1) in the TMLE procedure for estimating the bias from including external data, 2) as plug-in estimates for the relevant components of the efficient influence curve for the ATE parameter, used in variance estimation, and 3) to obtain initial estimates of $g_n(A|S \in \{1, s\}, W)$ and $Q_n(S \in \{1, s\}, A, W)$ for estimation set observations in the CV-TMLE ATE estimation step. Each of these steps will be discussed in further detail in the following sections. As described in Chapter 1, to estimate the bias from including external data in the analysis, we also must estimate: $E[Y|S \in \{1, s\}, S, A, W]$, $E[NCO|S \in \{1, s\}, A, W]$ if an NCO is available, and $P(S = 1|S \in \{1, s\}, A = 0, W)$. All regressions are estimated using the SuperLearner machine learning prediction algorithm of van der Laan et al.[61]. The SuperLearner uses cross-validation to select the best algorithm for a given prediction problem from a library of parametric or non-parametric candidate algorithms. Further details regarding specification of a SuperLearner may be found below.

2.1.2 Bias estimation

After training estimators of the outcome and treatment mechanism regressions, we may proceed

with bias estimation. The ES-CVTMLE defines the bias or causal gap from integrating external and RCT data for a given experiment as $\Psi_s^{\#}(P_0) = \Psi_s(P_0) - \tilde{\Psi}_s(P_0)$ [60] where

$$\tilde{\Psi}_s(P_0) = E_{W|S\in\{1,s\}}[E_0[Y|A=1,S=1,W] - E_0[Y|A=0,S=1,W]].$$

As described in Chapter 1, when only control participants are available in the external data, this bias term simplifies to

$$\Psi_s^{\#}(P_0) = E_{W|S\in\{1,s\}}[E_0[Y|A=0,S=1,W]] - E_{W|S\in\{1,s\}}[E_0[Y|A=0,S\in\{1,s\},W]].$$

Extending the method to consider contexts where active treatment is also available in the external data requires only a small change in bias estimation as

$$\Psi_s^{\#}(P_0) = \Psi_s(P_0) - \tilde{\Psi}_s(P_0)$$

$$= (E_{W|S\in\{1,s\}}[E_0[Y|A=1,S,S\in\{1,s\},W]] - E_{W|S\in\{1,s\}}[E_0[Y|A=0,S,S\in\{1,s\},W]])$$
$$-(E_{W|S\in\{1,s\}}[E_0[Y|A=1,S=1,W]] - E_{W|S\in\{1,s\}}[E_0[Y|A=0,S=1,W]]).$$

Note that when active treatment is available in the external data (meaning that $S=0$ does not deterministically imply $A=0$), we may now adjust for S as a baseline variable in all regressions, leading to estimation of $g_n(A|S\in\{1,s\},S,W) = P_n(A=a|S\in\{1,s\},S,W)$ and $Q_n(S\in\{1,s\},S,A,W) = E_n[Y|S\in\{1,s\},S,A,W]$ and adjustment for S in the final ATE estimate as well as the bias terms. For simplicity of exposition, however, we will use the notation for the case where active treatment is not available in the external data below. The first argument of the *ES.cvtmle* function handles these changes to the estimation procedure by setting **txinrwd = TRUE** if active treatment is available in the external or "real world" data and **txinrwd = FALSE** if only extra controls are considered. Regardless, we estimate $\Psi_s^{\#}(P_0)$ using a TMLE estimator, $\hat{\Psi}_s^{\#,TMLE}(P_n)$. Please see Chapter 1 for further details of the bias estimation process.

2.1.3 Second bias estimate: The ATE on a negative control outcome

Because the bias from including external data must be estimated, finite sample variability may lead either to overestimation of bias and exclusion of unbiased external data or underestimation of bias and inappropriate incorporation of biased external data. To improve our ability to distinguish whether the ATE estimator for a given experiment is biased, the ES-CVTMLE method utilizes a second, objective measure of bias in the experiment selection procedure: the estimated ATE of treatment on a negative control outcome. An ideal NCO requires careful consideration as this outcome should be affected by as many as possible of the unmeasured factors that bias the effect estimate for the treatment on the primary outcome, while not being affected by the treatment itself [24, 17]. Ideally, the magnitude of the effect of the unmeasured factors on the treatment-NCO association should be as similar as possible to the magnitude of the effect of the unmeasured factors on the treatment-primary outcome association so that the estimated ATE of treatment on the NCO is equal to the causal bias for the primary effect estimate. If not, more complex methods of evaluating bias using an NCO are required (e.g., [26, 25]).

Again using the g-computation formula [29], the statistical estimand for the ATE on an NCO from an experiment with $S\in\{1,s\}$ is given by

$$\Phi_s(P_0) = E_{W|S\in\{1,s\}}[E_0[NCO|W,A=1,S\in\{1,s\}] - E_0[NCO|W,A=0,S\in\{1,s\}]]$$

and is also estimated in the *EScvtmle* package using TMLE. While it is not necessary to have an NCO to use the *EScvtmle* package, we demonstrate in Chapter 1 that when the estimated ATE on the NCO ($\hat{\Phi}_s^{TMLE}(P_n)$) is added to the bias estimate $\hat{\Psi}_s^{\#,TMLE}(P_n)$ in the criterion used for experiment selection, this improves the probability that unbiased external data are included while maintaining close to nominal coverage across a range of potential magnitudes of external data bias, even when the NCO is only affected by some of the factors causing bias for the treatment-primary outcome ATE estimate. Note that the NCO should be chosen such that bias affects the treatment-NCO relationship in the same direction as the treatment-primary outcome relationship.

2.1.4 Variance estimation

Now that we have estimated the bias, the next step is variance estimation. The efficient influence curve (EIC) for $\Psi_s(P_0)$ is given by

$$D^*_{\Psi_s}(O) =$$
$$(\frac{I(S \in \{1,s\})}{P(S \in \{1,s\})})((\frac{I(A=1)}{g_0(A=1|S \in \{1,s\},W)} - \frac{I(A=0)}{g_0(A=0|S \in \{1,s\},W)})$$
$$(Y - Q_0(S \in \{1,s\},A,W))$$
$$+Q_0(S \in \{1,s\},1,W) - Q_0(S \in \{1,s\},0,W) - \Psi_s(P_0)).$$

$g_0(A|S \in \{1,s\},W)$ and $Q_0(S \in \{1,s\},A,W)$ are estimated using the SuperLearner, as described above. The variance of the TMLE ATE estimator for each experiment is estimated by the variance of the EIC divided by the sample size: $\frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n}$.

2.1.5 Experiment selection

In order to optimize the estimated bias-variance tradeoff of including external data, the potential experiment-selection criteria for the ES-CVTMLE are the bias squared plus variance or "**b2v**" selector,

$$s_n^\star = \underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + (\hat{\Psi}_s^{\#,TMLE}(P_n))^2$$

if no NCO is considered, or the "**nco bias**" selector,

$$s_n^{\star\star} = \underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + (\hat{\Psi}_s^{\#,TMLE}(P_n) + \hat{\Phi}_s^{TMLE}(P_n))^2$$

if an NCO is available. Using one of these criteria, the ES-CVTMLE selects one experiment (either RCT only or RCT combined with external data) in each fold.

2.1.6 Caution regarding large potential improvements in variance

Using the "**b2v**" criterion, the experiment-selector chooses the experiment with the lowest estimated mean squared error, yet that may not be the only objective. If an experiment is selected for which the bias is larger than the standard error of the estimator, this may lead to less than nominal

coverage. This could happen if bias is underestimated due to finite sample variability or in the following context, which is more likely when active treatment is available in the external data, leading to larger potential efficiency gains.

For the $s_n^\star$ (**b2v**) selector, bias is deterministically 0 for the RCT-only experiment. Even if the true bias and variance were used for selection, an experiment with $S \in \{1, s\}$ with larger bias than standard error could be selected if

$$\frac{\sigma^2_{D^*_{\Psi_s}}}{n} + (\Psi^{\#}_s(P_0))^2 < \frac{\sigma^2_{D^*_{\Psi_{s=1}}}}{n}$$

and

$$\frac{\sigma^2_{D^*_{\Psi_s}}}{n} < (\Psi^{\#}_s(P_0))^2$$

which would be possible if

$$\frac{\sigma^2_{D^*_{\Psi_s}}}{n} < \frac{1}{2}\frac{\sigma^2_{D^*_{\Psi_{s=1}}}}{n}.$$

If maintenance of nominal coverage is more important than minimization of mean squared error, then a user may limit the number of external participants considered such that the variance of the ATE estimator for the combined RCT-external data experiment would not be expected to have half the variance or less compared to the variance of the ATE estimator for the RCT-only experiment. As a result, selection of an experiment that would lead to an estimator with larger bias than standard error would only occur due to finite sample variability in bias estimation.

2.2.2 Estimation

The final goal of the ES-CVTMLE is to use estimation set data to estimate the ATE for the experiment selected in each experiment-selection set and to take the average of these ATE estimates across folds. The TMLE procedure for estimating the ATE requires using the outcome and treatment mechanism regressions trained in each experiment selection set for each experiment to predict values of $g_n(A|S \in \{1, s\}, W)$ and $Q_n(S \in \{1, s\}, A, W)$ for the corresponding estimation set observations for all folds. The initial $Q_n(S \in \{1, s\}, A, W)$ estimates are then "targeted" to solve the EIC for the ATE and to optimize the statistical bias-variance tradeoff for target parameter estimation [41].

Because the *EScvtmle* is meant to be used in contexts where the RCT of interest is small, we employ a pooled TMLE targeting step by pooling initial estimation set estimates of $g_n(A|S \in \{1, s\}, W)$ and $Q_n(S \in \{1, s\}, A, W)$ for each experiment across all folds. Then, for each experiment, we obtain an experiment-specific TMLE targeting coefficient, ϵ_s, by performing a logistic regression of the outcome (scaled to lie between 0 and 1 for continuous outcomes) on the clever covariate $H^*_{s,n}(S \in \{1, s\}, A, W) = \frac{2A-1}{g_n(A|S \in \{1,s\}, W)}$ with offset $logit(Q_n(S \in \{1, s\}, A, W))$.

In each estimation set, we then select the data for the experiment chosen in the corresponding experiment-selection set (using either $s_n^\star(v^c)$ or $s_n^{\star\star}(v^c)$), and use the experiment-specific coefficient to update initial fold-specific estimates $Q_{n,v^c}(S \in \{1, s_n^\star(v^c)\}, a, W)$ for $a \in \{0, 1\}$ as

$$Q^*_{n,v}(S \in \{1, s_n^\star(v^c)\}, a, W)$$

$$= logit^{-1}(logit(Q_{n,v^c}(S \in \{1, s_n^\star(v^c)\}, a, W)) + \epsilon_{s_n^\star(v^c)} * H^*_{s_n^\star(v^c),n}(S \in \{1, s_n^\star(v^c)\}, a, W))$$

where

$$H^*_{s^\star_n(v^c),n}(S \in \{1, s^\star_n(v^c)\}, 1, W) = \frac{1}{g_{n,v^c}(A = 1|S \in \{1, s^\star_n(v^c)\}, W)}$$

$$H^*_{s^\star_n(v^c),n}(S \in \{1, s^\star_n(v^c)\}, 0, W) = \frac{-1}{g_{n,v^c}(A = 0|S \in \{1, s^\star_n(v^c)\}, W)}.$$

The final ES-CVTMLE estimate is then an average of the fold-specific TMLE ATE estimates for the experiment selected in each fold:

$$\hat{\Psi}^{TMLE}_{s^*_n(v^c)}(Q^*_{n,v}) =$$
$$\frac{1}{n}\sum_{i=1}^{n} \frac{I(S_i \in \{1, s^\star_n(v^c)\}, \bar{V}_i = v)}{P_n(S \in \{1, s^\star_n(v^c)\}, \bar{V} = v)}[Q^*_{n,v}(S_i \in \{1, s^\star_n(v^c)\}, 1, W_i) - Q^*_{n,v}(S_i \in \{1, s^\star_n(v^c)\}, 0, W_i)]$$

yielding a final parameter estimate of $\psi_n = \frac{1}{V}\sum_{v=1}^{V} \hat{\Psi}^{TMLE}_{s^\star_n(v^c)}(Q^*_{n,v})$.

2.2.3 Confidence interval construction

Finally, we must construct 95% confidence intervals for the ES-CVTMLE estimate. If bias is large enough that external data is deterministically rejected, the ES-CVTMLE, like a standard CV-TMLE estimator for the ATE for a single experiment, converges to a normal distribution [14, 62, 15, 60]. For this reason, if only RCT data is selected in all folds, the ES-CVTMLE uses an influence-curve based variance estimate with $D^*_{\Psi_{s=1},n,v}$ estimated among estimation set RCT observations for each fold, $n_{s=1}$ equal to the sample size of RCT participants, and confidence intervals constructed as

$$\psi_n \pm 1.96 * (\frac{1}{V}\sum_{v=1}^{V} \frac{\hat{\sigma}^2_{D^*_{\Psi_{s=1},n,v}}}{n_{s=1}})^{1/2}.$$

When causal bias is of a magnitude small enough that external data is not deterministically rejected, the ES-CVTMLE converges to a mixture of normal distributions [60]. In the case where external data is included in at least one fold, confidence intervals are constructed via Monte Carlo sampling from the estimated limit distribution of the standardized estimator, as described in Chapter 1. Briefly, if we define the standardized TMLE estimators of the bias terms using experiment-selection set data and of the ATE using estimation set data for each experiment with $S \in \{1, s\}$ and each fold as

$$Z^{\#}_n(s, v^c) = \sqrt{n}(\hat{\Psi}^{\#,TMLE}_s(P_{n,v^c}) - \Psi^{\#}_s(P_0))$$
$$Z^{\#+\Phi}_n(s, v^c) = \sqrt{n}((\hat{\Psi}^{\#,TMLE}_s(P_{n,v^c}) + \hat{\Phi}^{TMLE}_s(P_{n,v^c})) - (\Psi^{\#}_s(P_0) + \Phi_s(P_0)))$$
$$Z_n(s, v) = \sqrt{n}(\hat{\Psi}^{TMLE}_s(P_{n,v}) - \Psi_s(P_0))$$

then across all experiments and across all folds, we may define vectors of these standardized estimators. For example, for the ATE estimator, we have

$$Z_n = (Z_n(s, v) : s = 0, 1, v = 1, ..., V) \sim Z = (Z(s, v) : s = 0, 1, v = 1, ..., V).$$

Then, as shown in Chapter 1, the stacked vector $\tilde{Z} = (Z^{\#}, Z) \sim N(\overrightarrow{0}, \tilde{\Sigma})$ or $\tilde{Z} = (Z^{\#+\Phi}, Z) \sim N(\overrightarrow{0}, \tilde{\Sigma})$, and $\tilde{\Sigma}$ is defined by the variance-covariance matrix of the efficient influence curves for each parameter in the stacked vector $\tilde{Z}$. Again using plug-in estimates for the relevant components of the EICs, we may estimate $\tilde{\Sigma}$ and obtain multiple samples from a mean zero multivariate normal distribution with this estimated covariance matrix. Finally, we obtain a single sample from the

estimated limit distribution of the standardized ES-CVTMLE starting with a single sample Z by 1) plugging $Z^{\#}$ or $Z^{\#+\Phi}$ and the bias and variance estimates described in Sections 2.2.1.2-4 into one of the standardized selectors

$$\bar{s}_n^{\star}(v^c) = \underset{s}{argmin}\ \hat{\sigma}^2_{D^*_{\Psi_{s,v^c}}} + (Z^{\#}(s,v^c) + \sqrt{n}\hat{\Psi}_s^{\#,TMLE}(P_{n,v^c}))^2$$

$$\bar{s}_n^{\star\star}(v^c) = \underset{s}{argmin}\ \hat{\sigma}^2_{D^*_{\Psi_{s,v^c}}} + (Z^{\#+\Phi}(s,v^c) + \sqrt{n}(\hat{\Psi}_s^{\#,TMLE}(P_{n,v^c}) + \hat{\Phi}_s^{TMLE}(P_{n,v^c})))^2$$

to select an experiment for each fold and 2) calculating $\frac{1}{V}\sum_{v=1}^{V}(Z(\bar{s}_n^{\star}(v^c),v))$ or $\frac{1}{V}\sum_{v=1}^{V}(Z(\bar{s}_n^{\star\star}(v^c),v))$ to obtain a single sample from the limit distribution. Repeating this process many times, with the p^{th} percentile of these samples denoted q^p, the final ES-CVTMLE confidence interval estimates are given by

$$\psi_n + (\frac{q^{0.025}}{\sqrt{n}}, \frac{q^{0.975}}{\sqrt{n}}).$$

Note that by including an estimate of the bias in this sampling process, the confidence interval width increases in response to larger estimated bias.

2.3 Implementation: The EScvtmle package

Next, we describe how the ES-CVTMLE methodology is implemented in the EScvtmle R package. To begin, first download the package from CRAN:

```
#install.packages("EScvtmle")
library(EScvtmle)
```

We will go over the specific functionality of the package in the sections below.

2.4 The ES.cvtmle function

The ES.cvtmle function is a wrapper for all internal functions necessary to a) use TMLE to estimate the bias from including external data in an ATE estimate, b) select the optimal experiment as RCT-only or RCT-combined with external data, c) use cross-validated targeted maximum likelihood estimation to estimate the statistical estimand that is the average of the ATE estimands for the experiment selected in each fold, $\psi_{n,0}$, and d) construct 95% confidence intervals as described above. The subsequent sections describe the arguments to the ES.cvtmle function.

2.4.1 Arguments specifying variables in the causal model

The first set of arguments to the ES.cvtmle function describe relevant variables in a causal model, including

- **data** The dataset
- **study** Character name of variable indicating study participation (e.g. "S"). This variable should take a value of 1 for the RCT and should take a value of 0 for the external data. Note that the code is currently set up only to handle two studies but may be expanded to handle multiple studies in the future.

- **covariates** Vector of character names of baseline covariates to be adjusted for (e.g. c("W1", "W2"))
- **treatment_var** Character name of treatment variable (e.g. "A")
- **treatment** Value of treatment variable that corresponds to the active treatment (e.g. "Drug-Name" or 1). All other values of the treatment variable are assumed to be control.
- **outcome** Character name of outcome variable (e.g. "Y"). If the outcome is a binary variable subject to censoring, censored observations should either be coded as NA or should be coded as 0, and a missingness indicator should be included (see parameter Delta below).
- **NCO** Character name of negative control outcome variable (e.g. "nco") or NULL if no NCO is available. If the NCO is a binary variable subject to censoring, censored observations should either be coded as NA or should be coded as 0, and a missingness indicator should be included (see parameter Delta_NCO below).
- **Delta** Character name of a variable that is 0 if an observation was censored (missing outcome) and 1 otherwise. Missing outcomes may also be coded as NA, in which case a Delta variable will be added internally. If no missing outcomes, set Delta=NULL.
- **Delta_NCO** Character name of a variable that is 0 if the value of NCO is missing and 1 otherwise. Missing NCOs may also be coded as NA, in which case a Delta_NCO variable will be added internally. If no missing NCO or no NCO, set Delta_NCO=NULL.

The current version of the EScvtmle package only handles baseline covariates and single time-point outcomes. We plan to extend the ES-CVTMLE methodology to handle time-to-event outcomes and other longitudinal data structures in future versions.

Based on these arguments, the internal *preprocess* function modifies the data by: 1) removing observations missing treatment information, 2) creating missingness indicators for the primary and negative control outcomes if these variables are subject to missingness and the arguments *Delta* and *Delta_NCO* are not specified, and 3) pruning the external data to avoid a violation of the positivity assumption.

This final step is necessary because, to satisfy the positivity assumption, we must have that $P(A = a|W = w, S \in \{1, s\}) > 0$ for all $a \in A$ and all w for which $P(W = w, S \in \{1, s\}) > 0$. Alternatively, if active treatment is available in the external data, we must have that $P(A = a|W = w, S = s) > 0$ for all $a \in A$ and all w, s for which $P(W = w, S = s) > 0$ and $S \in \{1, s\}$. When only extra controls are considered, if there are external data observations whose combination of baseline covariates are not represented in the RCT, then $P(A = 1|W = w, S \in \{1, s\}) = 0$ for those participants. In this setting, we can avoid a positivity violation by trimming the external data such that there are no observations with baseline covariate values outside the range of values represented in the RCT. This trimming is conducted internally by the *preprocess* function, but users should be aware that target populations for all considered experiments then change to target populations whose covariates are consistent with the RCT sample.

2.4.2 Handling of missing outcomes

Consistent with the other TMLE packages [45, 63], missing outcomes are handled in the *ES-cvtmle* package by modifying the target parameter to consider an intervention to prevent missingness as well as to assign treatment $A = a$. Practically, this leads to the following considerations. First, both the outcome regressions and the coefficient for the TMLE targeting step are estimated among observations whose outcomes were observed. Second, to account for informative missingness, the clever covariates used to estimate the coefficient for TMLE targeting of the initial outcome regression estimates for both bias terms and the final ATE estimate are modified by adding an indicator that the outcome was observed, Δ, to the numerator and the estimated probability of not being censored conditional on treatment and covariates to the denominator.

For example, ignoring for notational convenience the cross-validation fold structure and different possible experiments, the clever covariate for the ATE becomes: $H(A,W) = \frac{\Delta(2*A-1)}{P(\Delta=1|A,W)g(A|W)}$ [45]. For all observations, initial estimates of the outcome regression are targeted as described in Section 2.2.2 with modified covariates:

$$H(1,W) = \frac{1}{P(\Delta = 1|A = 1, W)g(A = 1|W)}$$

$$H(0,W) = \frac{-1}{P(\Delta = 1|A = 0, W)g(A = 0|W)}.$$

The final TMLE estimate of the ATE is an average of the difference in targeted estimates $Q^\star(1,W)$ and $Q^\star(0,W)$ for all observations (regardless of missingness) as

$$\hat{\Psi}^{TMLE} = \frac{1}{n}\sum_{i=1}^{n}(Q^\star(1,W) - Q^\star(0,W)).$$

2.4.3 Handling of cluster-randomized studies or repeated measures

Also consistent with the *tmle* package [45], the *EScvtmle* package allows the number of independent units to be different than the number of observations, as may be the case for cluster-randomized studies or studies with repeated measures. To accomplish this, the user may specify the **id** argument as the name of the variable describing independent units in the dataset. Variance estimation is modified by defining components of the relevant efficient influence curve vectors as the average of the EIC for observations with the same value of the id variable and defining the number of independent units as the number of unique **id**s. Participants with the same value of the **id** variable are also kept within the same validation set for both SuperLearner estimation and for the overall ES-CVTMLE cross-validation structure.

2.4.4 Estimation of regressions

All regressions, including the outcome regressions for the primary and negative control outcomes, the treatment and trial participation mechanism regressions, and the missingness mechanism regression are estimated using the SuperLearner ensemble machine learning prediction algorithm [61] implemented by the *SuperLearner R* package [66]. Specification of these SuperLearners is controlled by the following options, where RWD indicates "real world data" or any data external to the primary RCT of interest:

- **Q.SL.library** Candidate algorithms for SuperLearner estimation of outcome regressions

- **d.SL.library.RCT** Candidate algorithms for SuperLearner estimation of missingness mechanism for RCT-only
- **d.SL.library.RWD** Candidate algorithms for SuperLearner estimation of missingness mechanism for RCT+RWD
- **g.SL.library** Candidate algorithms for SuperLearner estimation of treatment mechanism for combined RCT/RWD analysis
- **Q.discreteSL** Should a discrete SuperLearner be used for estimation of outcome regressions? (TRUE/FALSE)
- **d.discreteSL** Should a discrete SuperLearner be used for estimation of missingness mechanism? (TRUE/FALSE)
- **g.discreteSL** Should a discrete SuperLearner be used for estimation of treatment mechanism? (TRUE/FALSE)
- **family** Either "binomial" for binary outcomes or "gaussian" for continuous outcomes
- **family_nco** Family for negative control outcome
- **cvControl** A list of parameters to control the cross-validation process for the SuperLearners. See ?SuperLearner for more details.
- **adjustnco** Should we adjust for the NCO as a proxy of bias in the estimation of the ATE of A on Y? (TRUE/FALSE). Default is FALSE.

Because rates of missingness may be very different for the RCT and the real world/external dataset, and flexible machine learning algorithms that may be appropriate for predicting more common outcomes may overfit for rare outcomes, we allow specification of different SuperLearner libraries for estimating the missingness mechanism in the RCT and the combined RCT-external datasets.

The **.discreteSL** options ask whether a discrete SuperLearner, which is the single best performing algorithm out of the library of candidate algorithms based on a pre-specified loss function, or an ensemble SuperLearner, which is the optimal weighted convex combination of all candidate algorithms, should be used to estimate the relevant regressions [61]. Fitting a discrete SuperLearner may be advantageous in contexts with sparse information, such as rare outcomes. Please see documentation for the *SuperLearner R* package [66] regarding other arguments that may be passed to the SuperLearner algorithm using the **cvControl** parameter. For an overview of how to specify a SuperLearner, including considerations such as the number of cross-validation folds and appropriate candidate algorithms for a given outcome type, sample size, and set of predictors, we recommend Phillips et al.[67].

Finally, the argument **adjustnco** asks whether to adjust for the negative control outcome when estimating the ATE of treatment on the outcome. Because the same unmeasured factors that affect the treatment-outcome relationship should also affect the NCO [24], the NCO may serve as a proxy of bias, and adjusting for the NCO may decrease the impact of the unmeasured common causes of treatment or study participation and the outcome on the effect estimate. However, adjusting for

the NCO could also worsen predictions for the outcome in some contexts, and so we recommend including a variable screening algorithm in the SuperLearner library when setting **adjustnco = TRUE**.

2.4.5 Parameters of the ES-CVTMLE

The remaining arguments to the *ES.cvtmle* function pertain to the experiment-selection, cross-validated TMLE, or confidence interval construction steps of the ES-CVTMLE method:

- **pRCT** The probability of randomization to treatment in the RCT
- **V** Number of cross-validation folds (default 10)
- **fluctuation** 'logistic' (default for binary and continuous outcomes), or 'linear' describing fluctuation for targeted maximum likelihood estimation (TMLE) updating.
- **comparisons** A list of the values of the study variable to include in the compared experiments. Set comparisons = list(c(1),c(1,0)) to compare RCT only to RCT + external data.
- **target.gwt** As in the tmle *R* package [45], if target.gwt is TRUE, the treatment mechanism is moved from the denominator of the clever covariate to the weight when fitting the coefficient for TMLE updating. Default TRUE.
- **bounds** Optional bounds for truncation of the denominator of the clever covariate. The default is $c(5/\sqrt{n}/log(n), 1)$.
- **MCsamp** Number of Monte Carlo samples from the estimated limit distribution to use to estimate quantile-based confidence intervals. Default 1000.

If the true, known probability of randomization to active treatment in the RCT is used for the treatment mechanism and there is no outcome missingness, then, due to the property of double-robustness, the TMLE for the ATE for the RCT is asymptotically unbiased even if the outcome regression is misspecified [41]. For this reason, we use the known randomization probability specified by argument **pRCT** for $P(A = 1|W, S = 1)$ when estimating the ATE for the RCT-only experiment.

The number of cross-validation folds for the overall ES-CVTMLE is specified by the argument **V**. Generally speaking, a larger number of cross-validation-folds is recommended in contexts of data sparsity, such as rare binary outcomes [67]. Once again, Phillips et al.[67] has a helpful overview guiding the choice of the number of cross-validation folds. Within the *ES.cvtmle* function, the *origami R* package [68] is used to maintain a consistent proportion of RCT observations and also of binary outcomes across experiment-selection and estimation sets.

The **fluctuation** argument defines whether a logistic or linear fluctuation should be used for estimating the coefficient for the TMLE targeting step. For a logistic fluctuation, a logistic regression of the outcome on the clever covariate, using the initial estimate of $logit(Q(S \in \{1, s\}, A, W))$ as an offset, is conducted. A logistic fluctuation is the default for both binary and continuous outcomes, where the continuous outcome is first scaled to lie between zero and one, because Gruber and van der Laan[69] showed that this method causes the TMLE fluctuation to respect the bounds of the observed data in the context of data sparsity. $Q^*_{n,v}$ estimates are returned to the original scale

for parameter estimation. Nonetheless, a linear fluctuation may also be performed for a continuous outcome by conducting a linear regression of the outcome on the clever covariate, using the initial estimate of $Q(S \in \{1, s\}, A, W)$ as an offset, by specifying **fluctuation = 'linear'**.

The **comparisons** argument controls which experiments are compared. In the current version of the *EScvtmle* package, **comparisons** must be specified as **list(c(1),c(1,0))** to compare the RCT-only experiment with $S = 1$ to the pooled RCT-external data experiment with $S \in \{0, 1\}$. In the future, we plan to extend the *EScvtmle* package to allow for comparison of more than two experiments.

As in the *tmle R* package [45], the **target.gwt** parameter moves the denominator of the clever covariate to the weights in the regression used to estimate the coefficient for the TMLE updating step. For example, for a TMLE to estimate the ATE, we could perform logistic regression of Y on $H^*_{s,n}(S \in \{1, s\}, A, W) = I(A = 1) - I(A = 0)$ with offset $logit(Q_n(S \in \{1, s\}, A, W))$ and weights $(\frac{I(A=1)}{g_n(A=1|S\in\{1,s\},W)} + \frac{I(A=0)}{g_n(A=0|S\in\{1,s\},W)})$ among observations with $S \in \{1, s\}$. We then update initial estimates of the conditional mean outcome as, for example, $Q^*_n(S \in \{1, s\}, 0, W) = logit^{-1}(logit(Q_n(S \in \{1, s\}, 0, W)) - \epsilon_s)$. Robins et al.[42], Rotnitzky et al.[43], and Tran et al.[44] suggest "targeting the weights" to improve stability when some observations have a low probability of having received one of the treatment options based on their covariates.

The **bounds** argument allows the user to specify bounds for truncation of the denominator of the clever covariate. Bounding further away from zero decreases variance at the expense of increasing bias [70]. Consistent with the *tmle R* package, we use default bounds of $c(5/\sqrt{n}/log(n), 1)$. For an in depth discussion of proper bounding of the propensity score, we refer the reader to Gruber et al.[70].

The final user-specified argument, **MCsamp**, determines the number of samples taken from the estimated limit distribution used for constructing confidence intervals, as described in Section 2.2.3. In Chapter 1, the default of 1000 resulted in accurate inference when unbiased external data were provided. Increasing **MCsamp** above 1000 may further increase the stability of confidence interval construction at the expense of increased computational burden.

2.5 ES.cvtmle function output

After the *ES.cvtmle* function has been specified and run, we obtain as output an object that is a list with the following components, where each component is provided for the bias squared plus variance (**"b2v"**) selector and also for the selector that includes an estimate of the ATE on an NCO in the bias term (**"ncobias"**) if an NCO is available:

- **ATE** Average treatment effect (ATE) point estimate for the ES-CVTMLE estimator.
- **foldATEs** ATE point estimates for each cross-validation fold.
- **g** A list of the same length as **comparisons** where each element of the list is a vector of the denominator of the covariate in front of the residual in the efficient influence curve for all observations in the experiment described by that element of **comparisons**. Values of g close to 0 or 1 may be used to diagnose practical near-positivity violations.
- **CI** Estimated 95% confidence intervals for the ATE estimates of the ES-CVTMLE estimator.
- **limitdistributionsample** Monte Carlo samples for the ATE estimates of the ES-CVTMLE estimator that are used to construct confidence intervals.

- **Var** Estimated variance of the ES-CVTMLE ATE estimate.
- **selected_byfold** Vector noting which experiment from the list of comparisons was selected in each cross-validation fold.
- **proportionselected** Proportion of all cross-validation folds in which external data were included in the analysis.

The goal of the *ES.cvtmle* function output is to provide information relevant for making the inner workings of the estimator more transparent. We recommend using the *print.Escvtmle* and *plot.Escvtmle* functions to summarize this information in a user-friendly format.

2.5.1 Print and plot functions

The *print.Escvtmle* function prints a summary of the ES-CVTMLE estimate, taking as its argument an object produced by the *ES.cvtmle* function. *print.Escvtmle* prints both the percent of folds in which external data were included and the point estimate and 95% confidence interval for the ES-CVTMLE ATE estimate. If an NCO is provided, the results using both the $s_n^\star$ ("**b2v**") and $s_n^{\star\star}$ ("**ncobias**") selector are provided.

The *plot.Escvtmle* function provides further insight into the results by printing two graphs: one of the fold-specific average treatment effect estimates for all cross-validation folds (color coded based on which experiment was selected in each fold), and the other of a histogram of the Monte Carlo samples that are used to construct confidence intervals. If an NCO is available, the plots are for the selector that includes the estimated average treatment effect on the NCO in the bias estimate. If not, the plots are for the selector that uses the estimated bias squared plus variance selector, without information from an NCO. The first plot may be useful for visualizing the similarity of the point estimates for the ATE across folds for the different selected experiments.

2.6 Real data example: WASH benefits data analysis

It is important to understand not only how to use the *EScvtmle* package but also how to consider whether one should use the ES-CVTMLE method or decide *a priori* to analyze randomized or observational data alone. To help users understand this decision-making process, we conduct an analysis of the WASH Benefits Bangladesh cluster randomized controlled trial (CRCT) dataset, available from `https://osf.io.wvyn4/`. Originally reported by Luby et al.[65], this trial evaluated the effect of several community-level interventions on child growth in Bangladesh. The sanitation intervention included construction of improved latrines as well as supplies and training to help families dispose of feces in a hygienic manner. Luby et al.[65] did not find evidence of an effect of this randomized sanitation intervention on child length-for-age Z-score (LAZ, a marker of childhood nutritional status).

To demonstrate the potential for unmeasured confounding in observational studies of WASH interventions, Arnold et al.[71] analyzed the control arm of the Bangladesh trial as an observational cohort. The authors examined the association between having an improved latrine at baseline and child LAZ, adjusting for baseline covariates, including age, sex, mother's age, mother's education, mother's height, level of food insecurity, number of people living in the compound, time required to reach the available water source, and ownership of various household assets. After adjustment for these variables, access to an improved latrine at baseline was associated with an increase in

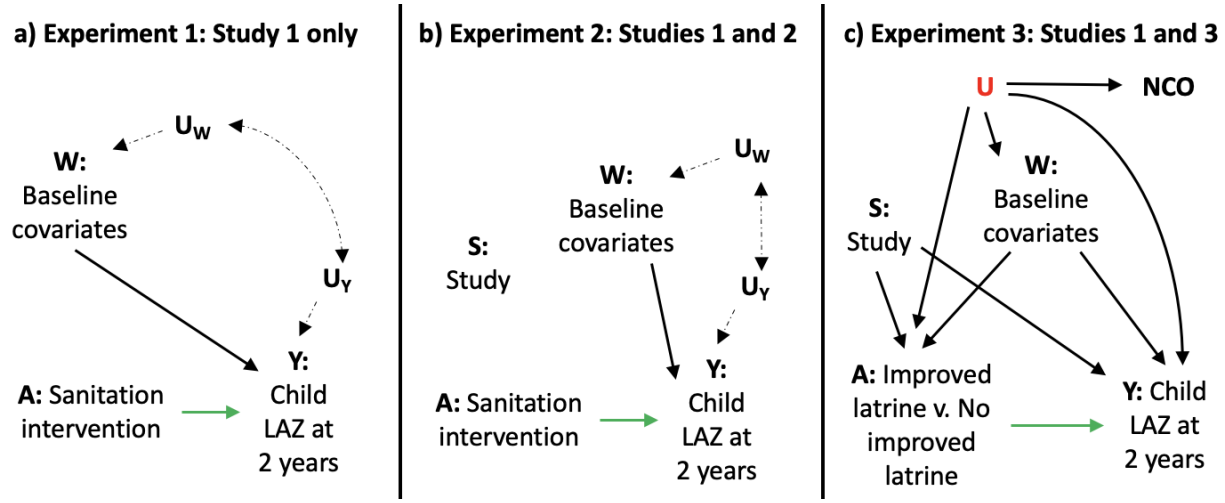
mean LAZ of 0.22 (95% CI 0.03-0.40) [71]. The analyses by Luby et al.[65] and Arnold et al.[71] thus provide an interesting example of a case where a large RCT and an observational analysis were conducted in the same population with different conclusions, likely due at least in part to bias from unmeasured confounding. Taking a subset of the full WASH Benefits Bangladesh CRCT to mimic an underpowered trial, we demonstrate the use of the *EScvtmle* package to include unbiased external data and exclude biased external data from our analysis.

In pursuance of this goal, we consider three experiments. The first includes only a small sample of the full RCT (study 1). The second considers adding additional RCT participants (study 2) to the study 1 sample as unbiased extra data. The third considers adding an observational cohort of control arm participants not included in study 1, where the intervention is defined as having a latrine with an improved water seal at baseline, consistent with the likely biased analysis of Arnold et al.[71]. To demonstrate how to think through whether a hybrid randomized-external data design should be considered, we describe this real data analysis using Petersen and van der Laan[27]'s causal roadmap.

2.6.1 Structural causal model

First, we describe what we know about the way the data for each experiment were generated. Figure 2.2 shows directed acyclic graphs for these three potential experiments. Because Experiment 1 includes only RCT data from a single "study", in this simple scenario, treatment is only affected by the randomization procedure. For Experiment 2, participants in studies 1 and 2 were randomly sampled from the overall RCT, so while there is now a study variable with values of 1 or 2, this variable is not affected by baseline characteristics and does not affect treatment or outcomes.

Figure 2.2: Directed acyclic graph for experiments 1-3



Our structural causal model for Experiment 3 is a bit more complicated. First, for study 3, Arnold et al.[71] define the intervention of interest as having (or not having) an improved latrine. In study 1, however, randomization to construction of an improved latrine, $A = 1$, was combined with the provision of other equipment and training. We will discuss whether it is wise to consider

combining studies 1 and 3 given this fact in Section 2.6.4 ("Identifiability") below. Wisdom aside, it is clear that which study a participant is in does affect the treatment they receive — because treatment is randomized in study 1 and determined by economic and other factors in study 3 — and may also affect their outcomes in other ways because the intervention in study 1 includes components beyond improved latrine construction. In this specific context, there are no unmeasured factors that affect study inclusion because study 1 is a random sample of the overall CRCT, and study 3 is a random sample of the remaining control arm participants. In study 3, however, there may be unmeasured common causes of the intervention, measured covariates, and outcome.

The last variable of interest is a negative control outcome. To select an appropriate NCO we must consider the likely unmeasured common causes of having an improved latrine at baseline and child LAZ in study 3, depicted by the U in Figure 2.2c. Although Arnold et al.[71] adjusted for some aspects of household wealth, we suspect that incomplete adjustment for socioeconomic status may have been a primary cause of the residual bias in this observational analysis. If designing a hybrid RCT-RWD study prospectively, an ideal NCO for this study would be an outcome measured at the same time as the primary outcome, on a similar scale, that is affected by socioeconomic status (SES) to a similar degree as the primary outcome. Nonetheless, in Chapter 1, we demonstrated that because the experiment-selector CV-TMLE relies on two different types of bias estimate, the NCO does not need to be perfect (in the sense of being affected by exactly the same factors that cause bias for the treatment-outcome relationship) in order to promote efficiency gains through the inclusion of unbiased external data while maintaining nominal or close to nominal coverage when only biased external data are available. We thus select one potential negative control variable from the available dataset as the number of household members less than or equal to 18 years old, because prior studies have shown this variable to be associated with SES in Bangladesh [72], but this number is unlikely to be affected by having an improved latrine. We scaled this variable to match the scale of the true outcome (length-for-age Z-score).

2.6.2 Causal question and target parameter

From these experiments, we would like to answer the following question; What would the difference in mean child LAZ at two years post-baseline be if all households in the Gazipur, Kishoreganj, Mymensingh, and Tangail districts of Bangladesh had an improved latrine (with or without provision of sanitation training and supplies) compared to if none of the households did? Using the same notation as in Section 2.2, our causal target parameter is the average of the ATE for the experiments selected in each cross-validation fold as:

$$\psi_0 = \frac{1}{V} \sum_{v=1}^{V} \Psi^F_{s^\star_n(v^c)}(P_{U,X})$$

2.6.3 Observed data

Now that we have described the experiments that generated the data and the question we would like to answer, let us load and examine the simplified subset of the WASH Benefits Bangladesh dataset included with the *EScvtmle* package.

```
data("wash", package = "EScvtmle")
colnames(wash)
```

For studies 1 and 2, the "intervention" variable is defined as being in the "Sanitation" arm of the CRCT ($A = 1$) versus the "Control" arm ($A = 0$), while for study 3 it is defined as having

an improved latrine at baseline ($A = 1$) or not ($A = 0$). The outcome of interest, "laz", is child length-for-age Z-score at two years post-baseline. "Nlt18scale" is the scaled number of household members less than or equal to 18 years old, which is used in this example as the negative control variable. The baseline covariates included in this sample dataset are age in days ("aged"), sex, mother's education level ("momedu"), and category of household food insecurity ("hfiacat"). This example dataset has no missing covariates or outcomes.

```
table(wash$study)
```

The different "studies" included in this example dataset were created as follows. First, study 1 was generated by taking a random sample of 150 "Sanitation" arm participants and 150 "Control" arm participants with complete information from the overall RCT. Study 2 was then created to mimic an unbiased external dataset by taking a second random sample of 150 "Sanitation" arm participants and 150 "Control" arm participants with complete information from the remaining RCT participants. Finally, to create a biased external dataset, we sampled from the "Control" arm participants who were not included in study 1, 150 participants who had improved latrines at baseline and 150 participants who did not have improved latrines at baseline. While the independent unit in this trial is actually the randomization block [65], for this example, we treat the individual as the independent unit in order to have a larger sample size for splitting the trial into different studies.

2.6.4 Identifiability and statistical estimand

The next question is whether we can estimate our causal target parameter from this observed data. Our main concern is whether the identification assumption of mean exchangeability is true. We know that "mean exchangeability in the trial" [33], indicating that for every $a \in A$, $E[Y^a|W, S = 1, A = a] = E[Y^a|W, S = 1]$, is true because of randomization in study 1. Because we have active treatment in the external data, we also need that $E[Y^a|W, S = s, A = a] = E[Y^a|W, S = s]$ is true for an external dataset with $S = s$. We know that this assumption is true for study 2 because treatment was also randomized, though we should note that we are really estimating the effect of the full sanitation intervention, not just the improved latrines, on the outcome in both Experiment 1 and Experiment 2. In study 3, however, we are concerned about residual unmeasured aspects of SES affecting both improved latrine ownership and child LAZ, though Arnold et al.[71] did attempt to adjust for some aspects of SES in their analysis. For these reasons, we have high confidence in our ability to estimate a causal effect from the data in Experiment 2 but are not certain whether we are able to estimate a causal effect from the data in Experiment 3.

We would be remiss to not briefly discuss the related assumption of "mean exchangeability in effect measure" [33] indicating that $E[Y^1 - Y^0|W, S = 1] = E[Y^1 - Y^0|W, S \in \{1, s\}]$. Because we have active treatment in the external data, if this assumption is not true, then we may still be able to identify the causal target parameter

$$\psi_0 = \frac{1}{V} \sum_{v=1}^{V} \Psi^F_{s^\star_n(v^c)}(P_{U,X})$$

from the observed data, but this parameter is then a weighted average of the causal ATE of A on Y in study 1 and the causal ATE of A on Y in combined studies 1 and 2 or 1 and 3, where the weighting depends on the number of folds in which each experiment is selected. Given that studies 1 and 2 are just random samples from the same RCT, it is clear that "mean exchangeability in effect measure" is true for Experiments 1 and 2. For Experiments 1 and 3, "mean exchangeability in effect measure" is unlikely to be true. Being in study 1 is likely to modify the effect of being assigned $A = 1$ on

outcomes because participants assigned to improved latrine construction also received training and other supplies, and $A = 0$ includes some participants who had an improved latrine at baseline. If we were to data-adaptively select Experiment 1 in some folds and Experiment 3 in others, then our causal target parameter would be a weighted average of the ATE of the randomized intervention and the intervention of baseline improved latrine ownership on the outcome, which does not have a straightforward causal interpretation. We thus only recommend attempting to integrate randomized and external data where the intervention is the same insofar as that it does not differ in ways that would be expected to modify the outcome.

This discussion of identification assumptions proves an important point. First, if we are very confident that the causal target parameter is identifiable (i.e. Experiment 2), then we may proceed with an analysis of the data from that experiment using standard methods and allowing greater efficiency gains than if the ES-CVTMLE were used to check for bias. Second, if we know that the causal target parameter is not identifiable (e.g., we know that we have not measured important common causes of treatment and the outcome in the external data), then we know that we cannot estimate a causal effect using data from that experiment, whether we use the ES-CVTMLE or not. If, however, we have a situation where we think the interventions in the RCT and observational contexts are very likely to be equivalent, and we believe it to be likely that we have succeeded in adjusting for relevant confounding factors like SES, we may wish to use the ES-CVTMLE to decide whether to analyze the RCT only or the RCT combined with external data. To demonstrate how the ES-CVTMLE can protect against bias if our identification assumptions are not true, we will proceed with estimating

$$\psi_{n,0} = \frac{1}{V} \sum_{v=1}^{V} \Psi_{s_n^\star(v^c)}(P_0)$$

considering Experiment 1 or Experiment 3 below.

2.6.5 Estimation and interpretation

Finally, we use the *EScvtmle* package to estimate our target parameter. For computational speed, we estimate all regressions with linear or logistic regression. First, we consider either Experiment 1 (study 1 alone) or Experiment 2 (studies 1 and 2), and re-code the study variable for the "external" data from study 2 as "study=0".

```
dat <- wash[which(wash$study %in% c(1,2)),]
dat$study[which(dat$study==2)] <- 0
```

Then, we run the *ES.cvtmle* function, setting the seed for reproducibility, and print a summary of the results.

```
set.seed(2022)

results_exp12 <- ES.cvtmle(txinrwd=TRUE,
                           data=dat, study="study",
                           covariates=c("aged", "sex", "momedu", "hfiacat"),
                           treatment_var="intervention", treatment=1,
                           outcome="laz", NCO="Nlt18scale",
                           Delta=NULL, Delta_NCO=NULL,
                           pRCT=0.5, V=10,
```

```
                        Q.SL.library=c("SL.glm"), g.SL.library=c("SL.glm"),
                        Q.discreteSL=TRUE, g.discreteSL=TRUE,
                        family="gaussian", family_nco="gaussian",
                        fluctuation = "logistic",
                        comparisons = list(c(1),c(1,0)),
                        adjustnco = FALSE, target.gwt = TRUE)

print.EScvtmle(results_exp12)
```

When considering Experiment 1 or Experiment 2, the ES-CVTMLE selected Experiment 2 (studies 1 and 2) in all folds. The estimated ATE of the WASH Benefits Bangladesh trial sanitation intervention on child length-for-age Z-score at two years post-baseline was 0.021 with a 95% confidence interval of (-0.198 to 0.259) for the **b2v** selector and a 95% confidence interval of (-0.176 to 0.231) for the **ncobias** selector. As expected, including the estimated ATE of the intervention on the NCO in the bias term improved the efficiency of the estimator. These results are consistent with the results of the overall CRCT in that they do not provide evidence of an effect of the sanitation intervention on child growth at two years.

If we had analyzed only the small study 1 sample using CV-TMLE from the *tmle* package [45], we would have had the following results:

```
dat <- wash[which(wash$study==1),]

set.seed(2022)

library(tmle)
results_study1 <- tmle(Y=dat$laz, A=dat$intervention,
                       W=subset(dat, select=c(aged, sex, momedu, hfiacat)),
                       Q.SL.library=c("SL.glm"), g.SL.library=c("SL.glm"),
                       Q.discreteSL=TRUE, g.discreteSL=TRUE,
                       family="gaussian", fluctuation="logistic",
                       cvQinit=TRUE, V=10,
                       prescreenW.g=FALSE, target.gwt=TRUE
                       )

results_study1$estimates$ATE
```

By integrating unbiased external data, the ES-CVTMLE results are more efficient, with confidence interval widths that are, relative to the study 1 CV-TMLE confidence interval widths, 0.91 for the **b2v** selector and 0.81 for the **ncobias** selector.

Next, we consider either Experiment 1 (study 1 alone) or Experiment 3 (studies 1 and 3).

```
dat <- wash[which(wash$study %in% c(1,3)),]
dat$study[which(dat$study==3)] <- 0

set.seed(2022)

results_exp13 <- ES.cvtmle(txinrwd=TRUE,
```

```
                        data=dat, study="study",
                        covariates=c("aged", "sex", "momedu", "hfiacat"),
                        treatment_var="intervention", treatment=1,
                        outcome="laz", NCO="Nlt18scale",
                        Delta=NULL, Delta_NCO=NULL,
                        pRCT=0.5, V=10,
                        Q.SL.library=c("SL.glm"), g.SL.library=c("SL.glm"),
                        Q.discreteSL=TRUE, g.discreteSL=TRUE,
                        family="gaussian", family_nco="gaussian",
                        fluctuation = "logistic", comparisons = list(c(1),c
                           ↪ (1,0)),
                        adjustnco = FALSE, target.gwt = TRUE)

print.EScvtmle(results_exp13)
```

When considering Experiment 1 or Experiment 3, the ES-CVTMLE with the **b2v** selector included study 3 in 30% of folds with an ATE estimate of 0.04 (95% CI -0.196 to 0.273), while the ES-CVTMLE with the **ncobias** selector only included study 3 in 20% of folds with an ATE estimate of 0.027 (95% CI -0.183 to 0.26). The negative control variable thus helped to exclude biased external data. Results in both cases were consistent with the RCT in not demonstrating an effect of having an improved latrine on child LAZ.

If, instead of running study 1 and using the ES-CVTMLE, we had simply collected observational data as in study 3 and used a standard CV-TMLE with no check for bias, we would have estimated the effect of improved latrines on child LAZ at two years using the following code:

```
dat <- wash[which(wash$study==3),]

set.seed(2022)

library(tmle)
results_study3 <- tmle(Y=dat$laz, A=dat$intervention,
                       W=subset(dat, select=c(aged, sex, momedu, hfiacat)),
                       Q.SL.library=c("SL.glm"), g.SL.library=c("SL.glm"),
                       Q.discreteSL=TRUE, g.discreteSL=TRUE,
                       family="gaussian", fluctuation="logistic",
                       cvQinit=TRUE, V=10,
                       prescreenW.g=FALSE, target.gwt=TRUE
                       )

results_study3$estimates$ATE
```

These results from our sample study 3 dataset are, unsurprisingly, similar to Arnold et al.[71]'s analysis of the full WASH Benefits Bangladesh control arm, erroneously suggesting that improved latrines significantly improve child growth in this context (ATE 0.32, 95% CI 0.10-0.55). Yet the ES-CVTMLE detects bias from integrating study 3 data with study 1 data and excludes study 3 in most folds, preventing this biased conclusion.

Through this analysis of the WASH Benefits Bangladesh CRCT [65], we demonstrate the ability of the *EScvtmle* package to distinguish external data that would bias a causal effect estimate from unbiased external data and to improve estimator efficiency when unbiased extra data are available. Because the efficiency gains and type 1 error control of any estimator that integrates RCT with external data will depend on characteristics of the data — such as sample size, outcome type, and availability of external data with both active treatment and control participants — we recommend that any statistical analysis plan using this estimator include outcome blind-simulations to evaluate estimator properties in the proposed context.

2.7 Future extensions

There are several limitations to the *EScvtmle* package that will be developed in future versions. First, the package only currently compares two potential experiments: 1) an RCT only or 2) an RCT combined with one external dataset. In the future, we plan to allow consideration of multiple potential external datasets which might consist of, for example, multiple different numbers of propensity-score matched external controls. Second, missing covariates are currently not handled by the package, requiring the user to impute or drop observations with baseline variable missingness. Third, the package currently only estimates the average treatment effect (or causal risk difference) at a single timepoint. Future work will expand the ES-CVTMLE methodology to different contexts such as time-to-event endpoints with multiple time-point interventions. The most up-to-date version of the package may be found at `https://github.com/Lauren-EylerDang/EScvtmle`. Please check back for updates.

2.8 Discussion

With the FDA's Framework for Real World Evidence [59] and Complex Innovative Trial Design Program [56] highlighting potential new uses of RWD to support the regulatory approval process, the number of hybrid RCT-external data studies is likely to increase over time. The *EScvtmle R* package implements the experiment-selector CV-TMLE methodology described in Chapter 1 for data-adaptively selecting and analyzing the optimal experiment defined as RCT only or RCT combined with external data. By explaining the functionality of the package and demonstrating the use of the package to distinguish biased from unbiased external data in a re-analysis of the WASH Benefits Bangladesh CRCT, we hope to make the ES-CVTMLE method accessible to statisticians designing and analyzing hybrid randomized-external data studies.

3 A Causal Roadmap for Generating High-Quality Real-World Evidence

3.1 Introduction

The 21st century has witnessed a dramatic increase in the quality, diversity, and availability of real-world healthcare data (RWD) in forms such as electronic health records and registry or claims databases[73]. In 2016, as part of a strategy to improve the efficiency of medical product development, the United States Congress passed the 21st Century Cures Act[74] that mandated the development of United States Food and Drug Administration (FDA) guidance on potential regulatory uses of real-world evidence (RWE) – defined as "clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD"[59]. Internationally, stakeholders including other regulatory agencies, industry, payers, academia, and patient groups have also increasingly endorsed the use of RWE to support regulatory decisions[75, 76]. Emerging sources of RWE under evaluation include pragmatic clinical trials, externally controlled trials or hybrid randomized-external data studies, and long-term follow-up studies[77, 78, 56].

There are multiple motivations for generating RWE. First, RWE has long been used in post-market safety surveillance to uncover the presence of rare adverse events not adequately evaluated by phase III randomized controlled trials (RCTs) for reasons including strict eligibility criteria, strict treatment protocols, limited patient numbers, and limited time on treatment and in follow-up[79]. Second, recent drug development efforts have more commonly targeted rare diseases or conditions without effective treatments[80]. RWD can be useful in such contexts when it is not practical to randomize enough participants to power a standard RCT or when there is an ethical imperative to minimize the number of patients assigned to the trial control arm[81, 4]. RWE was also highly valuable during the COVID-19 pandemic; observational studies reported timely evidence of vaccine booster effectiveness[82, 83], compared the effectiveness of different vaccines[84], and evaluated vaccine effectiveness during pregnancy[85].

Despite the many ways in which RWE may support policy or regulatory decision-making, the prospect of erroneous conclusions resulting from potentially biased effect estimates has led to appropriate caution when interpreting the results of RWE studies. One concern is data availability; data sources might not include all relevant information for causal estimation even in randomized studies that generate RWE. Another concern is lack of randomized treatment allocation in observational RWE. These issues create challenges for estimating a causal relationship outside of the "traditional" clinical trial space.

In an attempt to guide investigators towards better practices for RWE studies, there has been a blossoming of input from regulatory agencies, academia, and industry in the form of guidelines and frameworks addressing different stages of the process of evidence generation[59, 76, 86, 87,

88, 89, 90, 91, 92]. Yet incoming submissions to regulatory agencies lack standardization and consistent inclusion of all information that is relevant for evaluating the quality of evidence that may be produced by a given RWE study[89]. How, then, can we help investigators do a better job of estimating causal effects – and evaluating the plausibility of assumptions needed to estimate causal effects – based at least partially on RWD?

To help answer this question, the Forum on the Integration of Observational and Randomized Data (FIORD) meeting was held in Washington, D.C. November 17-18, 2022 to discuss perspectives from regulatory and federal medical research agencies, industry, academia, trialists, methodologists, and software developers. FIORD participants discussed their experiences with RWE guidance and best practices and identified necessary steps and priorities for broadening usage by investigators. Specifically, participants determined the need for a unifying structure to assist with specification of a complete analytic design for an RWE study, including both the statistical analysis plan and additional design elements relevant for optimizing and evaluating the quality of evidence produced.

The Causal Roadmap[41, 27, 93, 94, 95, 96, 97] (hereafter, the Roadmap) addresses this need because it is a general, adaptable framework for causal and statistical inference that is applicable to all studies that generate RWE, including studies with randomized treatment allocation and prospective and retrospective observational designs. It is consistent with existing guidance and makes the steps necessary for pre-specifying the analytic design of RWE studies explicit. The Roadmap includes steps of defining a study question and the target of estimation, defining the processes that generate data to answer that question, articulating the assumptions required to give results a causal interpretation, selecting appropriate statistical analyses, and pre-specifying sensitivity analyses. Following the Roadmap may lead to either 1) a fully specified analytic study design (including pre-specified analysis plan) that is sufficient to generate high-quality RWE; or, 2) an evidence-based decision that an RWE study to generate the required level of evidence is not currently feasible, with insights into what data would be needed to generate suitable RWE in the future.

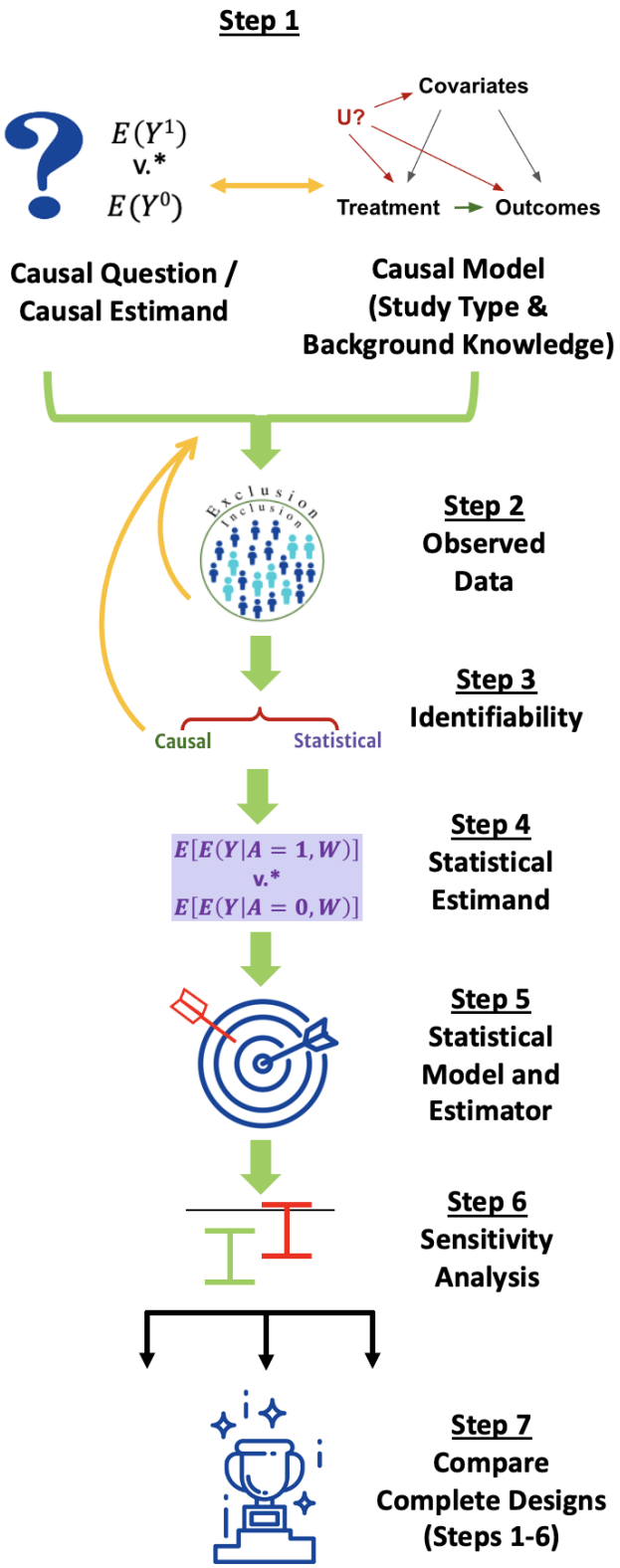
The goal of this chapter is to disseminate a Roadmap-based unifying framework for specifying analytic study designs for RWE generation to an audience of clinical and translational researchers. We provide an overview of the Roadmap, including a list of steps to consider when proposing studies that incorporate RWD.

3.2 Overview of the Causal Roadmap for clinical and translational scientists

We walk through the steps of the Roadmap, depicted in Figure 3.1, explaining their execution in general terms for simple scenarios, why they are important, and why multidisciplinary collaboration is valuable to accomplish each step. The structured approach outlined in Roadmap Steps 1-6 leads to specification of a complete analytic study design, which we define as including not only the type of study (e.g., randomized trial, observational cohort) but also elements of study design from the causal inference literature and the statistical analysis plan. The Roadmap does not cover all the steps necessary to write a protocol for running a prospective study, but instead specifies an explicit process for defining the study design itself, including information that is relevant for evaluating the quality of RWE that may be generated by that design. We suggest that following the Roadmap can help investigators generate high-quality RWE to answer questions that are important to patients, payers, regulators, and other stakeholders.

A century's worth of literature has contributed to the concepts described in the Roadmap. Several books explain nuances of these concepts[41, 98, 99, 100, 101, 102, 103]. This Chapter is not a comprehensive introduction, but rather aims to highlight steps that need to be considered to conduct high-quality causal inference and evidence generation.

Figure 3.1: The Causal Roadmap



* The contrast of interest may be additive (e.g., risk difference) or multiplicative (e.g., relative risk)

3.2.1 Step 1: Causal question, causal model, and causal estimand

As depicted in Figure 3.1, Step 1 involves defining a causal question, causal model, and the causal estimand that would answer that question. Formally, the causal model that describes relationships between key study variables would be defined before the causal estimand[27]. However, to facilitate explanation of these concepts, we start by using frameworks for specifying components of a causal estimand to also specify key elements of the causal model (Step 1a) before completing our causal model in Step 1b.

Step 1a: Define the causal question and causal estimand

Many causal questions start with the objective of estimating the effect of an exposure (e.g., a medication or intervention) on an outcome. Building on decades of research in the careful conduct of randomized and observational studies[103, 104, 105, 29, 106], both the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9(R1)[107] and Target Trial Emulation[86, 99, 108, 109] frameworks prompt investigators to define components of a causal question and estimand. The causal estimand is a mathematical quantity that answers the causal question (Table 3.1).

Table 3.1: Components of a Causal Question and Estimand per ICH E9(R1) [107] and Target Trial Emulation [86]

ICH E9(R1) attribute	Target Trial Emulation Protocol Component	Explanation	Related Notation in this Chapter
Population	Eligibility criteria	Inclusion and exclusion criteria, including dates of eligibility, for potential study participants	Measured baseline characteristics†: W
Treatment	Treatment strategies	The ideal hypothetical intervention(s) of interest in each arm of the target trial, including what treatment or exposure or intervention individuals would experience at study baseline and any post-baseline interventions, such as preventing censoring or requiring adherence for a specified duration.	Baseline intervention: A, Censoring††: C
	Follow-up period	The events that define the starting (e.g., randomization, prescription) and stopping (e.g., outcome, death) points for the observation period	

Variable or end point	Outcome	Outcome of interest, including the timepoint(s) at which the outcome will be evaluated	Outcome: Y
Population summary	Causal contrasts of interest	Causal Estimand†††: e.g., average treatment effect, relative risk, average treatment effect within prespecified subgroups	See below

† Baseline participant characteristics can include additional variables not used to define eligibility criteria. Baseline variables do not completely characterize the population, but for simplicity, we only consider measured baseline characteristics in the notation below.
†† In the current chapter, we focus as an example on interventions on baseline treatment and postbaseline censoring. However, the approach represented extends naturally to treatment strategies that incorporate additional postbaseline interventions, (see e.g., [110, 95])
††† A mathematical quantity that is a function of potential outcomes (see below).

An example of a question guided by these attributes might be: How would the risk of disease progression by 2 years have differed if all individuals who met eligibility criteria had experienced treatment strategy A=1 (e.g., drug under investigation) versus treatment strategy A=0 (e.g., active comparator) and no one dropped out of the study (C=0)? The best (albeit impossible!) way to answer this question would be to evaluate both the potential outcomes[28, 111] individuals would have had if they had experienced treatment strategy A=1 and not been censored ($Y^{a=1,c=0}$) and the potential outcomes individuals would have had if they had experienced treatment strategy A=0 and not been censored ($Y^{a=0,c=0}$).

A formal structural causal model would help us describe the causal pathways that generate these potential outcomes[112]. For now, we simply consider that, if we were able to observe both potential outcomes for all members of our target population, then the answer to our question would be given by the causal risk difference (or "Average Treatment Effect"),

$$\Psi^\star = P(Y^{a=1,c=0} = 1) - P(Y^{a=0,c=0} = 1).$$

This mathematical quantity that is a function of potential outcomes is called a causal estimand. Table 3.1 lists other examples of causal estimands.

Importance: Even though we can only observe at most one potential outcome for each individual [113], and even though it is not possible to guarantee complete follow-up in a real trial, precise definition of the causal question and estimand based on the treatment strategies defined in Table 3.1 is crucial for specifying a study design and analysis plan to provide the best possible effect estimate. Ultimately, we need to evaluate a mathematical expression that translates the available data into a number (e.g., a 5% decrease in risk of disease progression). To assess whether that number provides an answer to our causal question, we must first define mathematically what we aim to estimate.

Build a Multidisciplinary Collaboration: If you are not certain how to translate your research question into a causal estimand, collaborate with an expert in causal inference.

Step 1b: Specify a causal model describing how data have been or will be generated

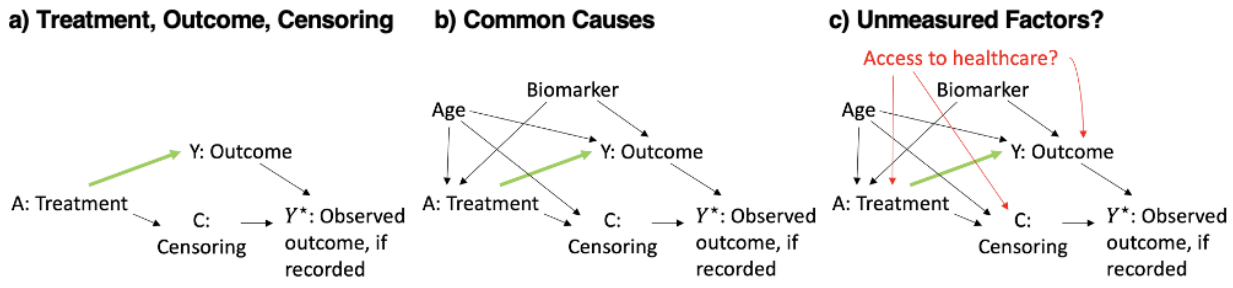
Next, we consider what we know (and don't know) about the real-life processes that will generate – or that have already generated – data to answer this question. First, we consider the type of study (e.g., pragmatic RCT, retrospective cohort study). Then, we consider what factors affect the variables that are part of our treatment strategies – found in Table 3.1 and referred to as intervention variables below – and the outcome in our proposed study.

This background knowledge comprises the causal model[112]. We specified some key variables in our causal model in Step 1a (in Table 3.1 and our potential outcomes). Now, we add additional detail to our causal model by describing potential causal relationships between these and other important variables. Multiple tools and frameworks can help elicit this information, but conceptual models and causal graphs, such as directed acyclic graphs or single world intervention graphs, are some of the most common[29, 114, 115, 116, 117].

Figure 3.2 gives a simple example of causal graph construction, starting with writing down all intervention and outcome variables. When some outcomes are missing, we don't observe the outcome, Y, for all participants. Instead, we observe $Y^{\star}$, which is equal to the actual outcome if it was observed and is missing otherwise (Figure 3.2a). Arrows denote possible effects of one variable on another.

Then, we attempt to write down factors that might influence these variables. Figure 3.2b shows two examples (age and a biomarker), though real causal graphs generally include many more variables. In a classic randomized trial, only the randomization procedure affects baseline treatment assignment, whereas in an observational study (depicted in Figure 3.2), participant characteristics affect the baseline treatment. Next, we consider factors that are unmeasured or difficult to measure that might influence treatment, outcomes, or censoring. Figure 3.2c shows access to healthcare as an example.

Figure 3.2: Basic Process for Generating a Causal Graph



Caption: $Y^{\star}$ is equal to the actual outcome value if it was observed and is missing otherwise.

Causal graphs can become much more complicated, especially when working with longitudinal data[29], using proxies for unmeasured variables[118], or combining different data sources[6] (as

demonstrated in the case study of Semaglutide and Cardiovascular Outcomes). A carefully constructed causal graph should also demonstrate issues such as competing risks, intercurrent events, or measurement error[99, 119].

Importance: Considering which factors may affect intervention variables and outcomes helps to determine whether we can answer our question based on existing data or data that we will collect. The final graph should be our best honest judgement based on available evidence and incorporating remaining uncertainty[99].

Build a Multidisciplinary Collaboration: If questions remain about some aspect of this model, such as how physicians decide to prescribe a medication in different practice settings, obtain input from clinicians or other relevant collaborators before moving on.

Stop! Return to Step 1a. Do you need to modify your causal question and estimand based on Step 1b?

After writing down our causal model, we sometimes need to change our question[120]. For example, we may have realized that an intercurrent event (such as death) prevents us from observing the outcome for some individuals. As suggested by ICH E9(R1), we could modify the question to consider the effect on a composite outcome of the original Y or death[107]. ICH E9(R1)[107] discusses other intercurrent events and potential modifications to the estimand.

3.2.2 Step 2: Describe the observed data

The causal model from Step 1b lets us specify what we know about the real-world processes that generate our observed data. This model can inform what data we collect in a prospective study or help to determine whether existing data sources include relevant information. Next, we describe the actual data we will observe.

Specific questions to address regarding the observed data include the following: How are the relevant exposures, outcomes, and covariates, including those defining eligibility criteria, measured in the observed data? Are they measured differently (including different monitoring protocols) in different data sources or at different timepoints? Are we able to measure all variables that are important common causes of the intervention variables and the outcome? Is the definition of time zero in the data consistent with the causal question[108]?

Importance: After considering these questions, we may need to modify Step 1. For example, if we realize that the data we are able to observe only include patients seen at tertiary care facilities, we may need to change the question (Step 1a) to ask about the difference in the risk of disease progression by two years if all individuals meeting our eligibility criteria and receiving care at a tertiary facility received one intervention or the other. Knowledge about factors that affect how variables are measured and whether they are missing should be incorporated in the causal model (Step 1b). Completing this step also helps investigators assess whether the data are fit-for-use[59] and whether we are able to estimate a causal effect from the observed data (discussed in Step 3).

Build a Multidisciplinary Collaboration: If you are unsure about the way variables are measured in relation to underlying medical concepts or in relation to a particular care setting, collaborate

with a clinician or clinical informaticist. If you are unsure of how to match baseline time zero in your observed data with the follow-up period in your causal question, collaborate with a statistician.

3.2.3 Step 3: Assess Identifiability: Can the proposed study provide an answer to our causal question?

In Step 3, we ask whether the data we do observe (Step 2), together with our knowledge about how these data are generated (Step 1b), are sufficient to let us answer our causal question (Step 1a). As described in Step 1a, we cannot directly estimate our causal estimand (which is a function of counterfactual outcomes). Instead, we will evaluate a function of the observed data (called a statistical estimand, described in Step 4). The difference between the true values of the statistical and causal estimands is sometimes referred to as the causal gap[94]. If there is a causal gap, even a perfect estimate of the statistical estimand would not provide an answer to our causal question.

While we can never be certain of the size of the causal gap for studies incorporating RWD and even for many questions using data from traditional RCTs, we must use our background knowledge to provide an honest appraisal. Causal identification assumptions help us to explicitly state what must be true in order to conclude that the causal gap is zero and that we are thus able to estimate a causal effect using the proposed data. Table 3.2 lists two common identification assumptions to consider for most cases with informal explanations of their meaning. Exchangeability, in particular, can also be framed in terms of causal graphs[112]. In some cases, further assumptions may be necessary. Hernán and Robins (2020)[99], among others, provide in-depth discussions of identification assumptions. The case studies associated with this paper demonstrate the evaluation of these assumptions.

Table 3.2: Common Identification Assumptions

Assumption	**Basic Explanation of Meaning**
Exchangeability[†]	This assumption is generally true if there are no unmeasured common causes of variables that are part of the treatment strategies (Table 3.1: e.g., baseline or postbaseline treatment(s), censoring) and the outcome (informally, if there is no unmeasured confounding).
Positivity	This assumption is true if, for every possible combination of measured confounding variables, individuals with those characteristics have a positive probability of following any of the treatment strategies of interest.

[†] Full exchangeability is generally not required if weaker conditions (e.g., mean exchangeability, sequential conditional exchangeability, or others) hold[99].

Importance: Considering and documenting the plausibility of the causal identification assumptions helps to determine whether steps can be taken to decrease the potential magnitude of the causal gap. If we conclude that these assumptions are unlikely to be satisfied, then we should consider modifications to Steps 1-2. We may need to limit the target population to those who have

a chance of receiving the intervention or evaluate the effect of a more realistic treatment rule to improve the plausibility of the positivity assumption[31, 121]. We may need to measure more of the common causes depicted in our causal graph or modify the question to improve the plausibility of the exchangeability assumption[122]. If multiple study designs are feasible, Step 3 can help us to consider which study design is based on more reasonable assumptions[123].

If we know that a key variable affecting treatment and outcomes or censoring and outcomes is not measured, then we generally cannot identify a causal effect from the observed data without measuring that variable or making additional assumptions[86, 99, 104]. For this and other reasons, many studies analyzing RWD appropriately report statistical associations and not causal effects, though sensitivity analyses (Step 6) may still help to evaluate whether a causal effect exists[124, 125]. Nonetheless, if a retrospective study was initially proposed but the causal identification assumptions are highly implausible and cannot be improved using existing data, then investigators should consider prospective data collection to better evaluate the effect of interest.

In general, it would be unreasonable to expect that all causal identification assumptions would be exactly true in RWE studies, or even in many traditional RCTs due to issues such as informative missingness[99]. Yet careful documentation of Steps 1-3 in the pre-specified analysis plan and in the study report helps not only the investigator but also regulators, clinicians, and other stakeholders to evaluate the quality of evidence generated by the study about the causal effect of interest. Step 3 helps us to specify a study with the smallest causal gap possible. Sensitivity analyses, discussed in Step 6, help to quantify a reasonable range for the causal gap, further aiding in the interpretation of RWE study results.

Build a Multidisciplinary Collaboration: An expert in causal inference can help to formally evaluate all causal identification assumptions. The exchangeability assumption can become quite complicated if there are multiple intervention variables[29, 126]. In such cases, graphical criteria may be used to determine visually from a causal graph whether sufficient variables have been measured to satisfy the exchangeability assumption[29, 112, 127]. Software programs can also facilitate this process[128, 129].

3.2.4 Step 4: Define the statistical estimand

If, after assessing identifiability, we decide to proceed with our study, we aim to define a statistical estimand that is as close as possible to the causal estimand of interest. Recall our causal risk difference for a single time-point intervention and outcome:

$$\Psi^\star = P(Y^{a=1,c=0} = 1) - P(Y^{a=0,c=0} = 1).$$

In a simple case where potential confounders – denoted W – are only measured at baseline, then the statistical estimand that is equivalent to the causal effect if all identification assumptions are true is given by

$$\Psi = E_W(P[Y^\star|C = 0, A = 1, W] - P[Y^\star|C = 0, A = 0, W]).$$

In words, we have re-written our causal question (which is defined based on potential outcomes that we cannot simultaneously observe) in terms of a quantity that we can estimate with our data: the average (for our target population) of the difference in risk of our observed outcome associated

with the different treatment strategies, adjusted for measured confounders.

Importance: The traditional practice of defining the statistical estimand as a coefficient in a regression model has several downsides, even if the model is correct (a questionable assumption discussed below)[41]. This approach starts with a tool (e.g., a regression model) and then asks what problem it can solve, rather than starting with a problem and choosing the best tool[130]. For example, the hazard ratio may be estimated based on a coefficient in a Cox regression but does not correspond to a clearly defined causal effect[131, 132, 133]. Instead, the Roadmap guides us to choose a statistical estimand that is as close as possible to the causal estimand. We thus specify a well-defined quantity that can be estimated from the observed data and that is directly linked to the causal question.

Build a Multidisciplinary Collaboration: Defining a statistical estimand that would be equivalent to the causal effect of interest under identification assumptions is more challenging when there are post-baseline variables that are affected by the exposure and that, in turn, affect both the outcome and subsequent intervention variables[29]. This situation is common in studies where the exposure is measured at multiple time-points. In such a situation, statistician collaborators can help to define the statistical estimand using approaches such as the longitudinal g-computation formula[29].

3.2.5 Step 5: Choose a statistical model and estimator that respects available knowledge and uncertainty based on statistical properties

The next step is to define a statistical model (formally, the set of possible data distributions) and to choose a statistical estimator. The statistical model should be compatible with the causal model (Step 1b). For example, knowledge that treatment will be randomized (design knowledge that we described in our causal model) implies balance in baseline characteristics across the two arms (with slight differences due to chance in a specific study sample). We could also incorporate knowledge that a continuous outcome falls within a known range or that a dose-response curve is monotonic (e.g., based on prior biological data) into our statistical model. A good statistical model summarizes such statistical knowledge about the form of the relationships between observed variables that is supported by available evidence without adding any unsubstantiated assumptions (such as linearity, or absence of interactions); models of this type are often referred to as semi- or non-parametric or simply realistic statistical models[41].

Given a statistical model, the choice of estimator should be based on pre-specified statistical performance benchmarks that evaluate how well it is likely to perform in estimating the statistical estimand[41]. Examples include type 1 error control, 95% confidence interval (CI) coverage, statistical bias, and precision. Statistical bias refers to how far the average estimate across many samples would be from the true value of the statistical estimand. An estimator must be flexible enough to perform well even when we do not know the form of the association between variables in our dataset, and it must be fully pre-specified[41].

Most available estimators rely on estimating an outcome regression (i.e., the expected value of the outcome given the treatment and values of confounders), a propensity score (i.e., the probability of receiving a treatment or intervention given the measured confounders), or both. Without knowing the form of these functions, we do not know a priori whether they are more likely to be accurately modeled with a parametric regression or a flexible machine learning algorithm allowing for

non-linearities and interactions between variables[41, 130, 61]. The traditional practice of defaulting to a parametric regression as the statistical estimator imposes additional untestable statistical assumptions, even though they are not necessary. Fortunately, estimators exist that allow for full pre-specification of all machine learning and parametric approaches used, data-adaptive selection (e.g., cross-validation) of the algorithm(s) that perform best for a given dataset, and theoretically-sound 95% confidence interval construction (leading to proper coverage under reasonable conditions)[41].

Importance: Effect estimates that are based on incorrectly specified models – such as a main terms linear regression when there is truly non-linearity or interactions between variables – are biased, and that bias does not get smaller as sample size increases[41]. This bias may result in inaccurate conclusions. We aim to choose an estimator that not only has minimal bias but also is efficient – thereby producing 95% confidence intervals that are accurate but as narrow as possible – to make maximal use of the data[41].

If, after consideration of the statistical assumptions and properties of the estimators, multiple estimators are considered, then the bias, variance, and 95% CI coverage of all estimators should be compared using outcome-blind simulations that mimic the true proposed experiment as closely as possible[134]. We use outcome-blind to mean that the simulations are conducted without information on the observed treatment-outcome association; such simulations may utilize other information from the collected data (if available), such as data on baseline covariates, treatment, and censoring, to approximate the real experiment[134]. Simulations conducted before data collection may use a range of plausible values for these study characteristics[135]. As recommended by ICH E9(R1), simulations should also be conducted for cases involving plausible violations of the statistical assumptions of the estimators[107]. Examples of such violations include non-linearity for linear models or inaccurate prior distributions for Bayesian parameters.

Build a Multidisciplinary Collaboration: Statistician collaborators can help to pre-specify an estimator with the statistical properties described above. Resources are increasingly available to assist with pre-specification of statistical analysis plans (SAPs) based on state-of-the-art estimation approaches. For example, Gruber et al. (2022)[136] provide a detailed description of how to pre-specify a SAP using targeted minimum loss-based estimation (TMLE)[40] and super learning[61], a combined approach that integrates machine learning to minimize the chance that statistical modeling assumptions are violated[41].

3.2.6 Step 6: Specify a procedure for sensitivity analysis

Sensitivity analyses in Step 6 attempt to quantify how the estimated results (Step 5) would change if the untestable causal identification assumptions from Step 3 were violated[99, 124, 137, 138, 139]. In contrast, the simulations in Step 5 consider bias due to violations of testable statistical assumptions, which ICH E9(R1) considers as a different form of sensitivity analysis[107]. One mechanism of conducting a causal sensitivity analysis in Step 6 is to consider the potential magnitude and direction of the causal gap; this process requires subject matter expertise and review of prior evidence[124, 138, 139, 140]. Sensitivity analysis also allows for construction of confidence intervals that account for plausible values of the causal gap[94, 124, 138, 139, 140]. Alternatively, investigators may assess for causal bias using negative control variables, discussed in detail by Lipsitch et al. (2010)[24] and Shi et al. (2020)[25].

The specifics of these methods – and alternate approaches – are beyond the scope of this chap-

ter, but note that the method of sensitivity analysis should be pre-specified prior to estimating the effect of interest[141]. This process avoids the bias that might occur if experts know the value of the estimate before defining the procedure they will use to decide whether a given shift in that estimate due to bias is reasonable[138].

Importance: The process of using prior evidence to reason about likely values of the causal gap helps investigators to assess the plausibility that the bias due to a violation of identification assumptions could be large enough that the observed effect is negated[94, 124, 125, 142]. While the exact magnitude of the causal effect may still not be identified due to known issues such as the potential for residual confounding, if an estimated effect is large enough, we may still obtain credible evidence that an effect exists[125, 143]; this was the case in Cornfield et al. (1959)'s frequently-cited assessment of the effect of smoking on lung cancer[144]. Conversely, if the anticipated effect size is small and the plausible range of the causal gap is large, the proposed study may not be able to provide actionable information. Considering these tradeoffs can help investigators to decide whether to pursue a given RWE study or to consider alternate designs that are more likely to provide high-quality evidence of whether a causal effect exists[125, 145].

Build a Multidisciplinary Collaboration: If multiple correlated sources of bias are likely, more complex methods of evaluating a plausible range for the causal gap – and collaboration with investigators familiar with these methods – may be required[138].

3.2.7 Step 7: Compare alternative complete analytic study designs

Roadmap Steps 1-6 help us to specify a complete analytic study design, including the causal question and estimand, type of study and additional knowledge about how the data are generated, specifics of the data sources that will be collected and/or analyzed, assumptions that the study relies on to evaluate a causal effect, statistical estimand, statistical estimator, and procedure for sensitivity analysis. The type of study described by this analytic design could fall anywhere on the spectrum from a traditional RCT to a fully observational analysis. In cases when it is not possible to conduct a traditional RCT due to logistical or ethical reasons – or when RCT results would not be available in time to provide actionable information – the value of RWE studies is clear despite the possibility of a causal gap[99]. If an RCT is feasible, baseline randomization of an intervention (as part of either a traditional or pragmatic RCT[146]) still generally affords a higher degree of certainty that the estimated effect is causal compared to analysis of non-randomized data. Yet sometimes, it is feasible to consider multiple different observational and/or randomized designs – each with different potential benefits and downsides.

Consider a situation in which there is some evidence for a favorable risk-benefit profile of a previously studied intervention based on prior data, but those data are by themselves insufficient for regulatory approval for a secondary indication or for clear modification of treatment guidelines. In this context, it is possible that conducting a well-designed RWE study or hybrid RCT-RWD study as opposed to a traditional RCT alone will shorten the time to a definitive conclusion, decrease the time patients are exposed to an inferior product, or provide other quantifiable benefits to patients while still providing acceptable control of type I and II errors[147, 3, 148]. Yet other times, a proposed RWE design may be inferior to alternative options, or one design may not be clearly superior to another. When multiple study designs are considered, outcome-blind simulations consistent with our description of Steps 1-6 can help to compare not only type 1 error and power, but also metrics

quantifying how the proposed designs will modify the medical product development process[147]. The case study of Semaglutide and Cardiovascular Outcomes in Chapter 4 demonstrates how to compare study designs that are based on Roadmap Steps 1-6.

Importance: A simulated comparison is not always necessary; one study design may be clearly superior to another. Yet often there are tradeoffs between studies with different specifications of Roadmap Steps 1-6. For example, in some contexts, we may consider augmenting an RCT with external data. When comparing the RCT design to the augmented RCT design, there may be a tradeoff between a) the probability of correctly stopping the study early when appropriate external controls are available and b) the worst-case type 1 error that would be expected if inappropriate external controls are considered[148]. Another example would be the tradeoff between the potential magnitudes of the causal gap when different assumptions are violated to varying degrees for studies relying on alternate sets of causal identification assumptions[123]. Simulated quantification of these tradeoffs using pre-specified benchmarks can help investigators to make design choices transparent[149].

Build a Multidisciplinary Collaboration: Factors to consider when comparing different analytic designs include the expected magnitude of benefit based on prior data and the quality of that data[81], the plausible bounds on the causal gap for a given RWE study, the treatments that are currently available[81], and preferences regarding tradeoffs between design characteristics such as type I versus type II error control[149]. Because these tradeoffs will be context-dependent[81, 149], collaboration with patient groups and discussion with regulatory agencies is often valuable when choosing a study design from multiple potential options.

3.3 A list of Roadmap steps for specifying a complete analytic study design

Table 3.3 provides a list of considerations to assist investigators in completing and documenting all steps of the Roadmap. Complete reporting of RWE study results should include all pre-specified Roadmap steps, though information supporting decisions in the final design and analysis plan, such as causal graphs or simulations, may be included as supplementary material. Note that all steps should be pre-specified before conducting the study.

Table 3.3: Steps for Specifying a Complete Analytic Study Design Using the Roadmap

1a	Specify the causal question and estimand. • ICH E9(R1) attributes: Population, treatment, variable or end point, population summary [107] • TTE Protocol Components: Eligibility criteria, treatment strategies, follow-up period, outcome, causal contrasts of interest [99]

1b	Specify the causal model (background knowledge about the proposed study). • Specify the type of study (e.g., traditional RCT, retrospective cohort) • Document whether censoring, competing risks, or other intercurrent events occurred and factors that may have affected them. Adjust the question as needed.
2	Define the observed data that has been or will be collected. • Document how inclusion/exclusion criteria, treatment variables, outcome(s), and other relevant variables are measured, how time zero is defined, and important differences between data sources.
3	Assess identifiability of the causal estimand from the observed data. • Explicitly state the assumptions required for identification, and evaluate their plausibility. • Consider modifications to Steps 1-2 to minimize the causal gap. – If a retrospective study had been planned but identification assumptions are highly implausible, consider primary data collection or linkage of data from different sources as necessary to ensure relevant information capture for the causal question and estimand.
4	Define the statistical estimand.
5	Specify the statistical model, estimator, and method of confidence interval construction. • List the assumptions the proposed estimator and method of confidence interval construction rely upon. • Describe the expected statistical bias and variance of the estimator under plausible conditions. • If multiple estimators are considered, compare them with outcome-blind simulations based on: – statistical bias, variance, 95% CI coverage of the statistical estimand, type 1 error, and power – with plausible violations of model assumptions.

6	Specify the sensitivity analyses. • Document the method for defining plausible bounds for the causal gap and/or methods for estimation of the causal gap (e.g., based on negative controls). • Provide confidence intervals for the causal effect of interest under the hypothesized size of the causal gap, across the full range of plausible causal gaps.
7	Compare feasible complete analytic designs (Steps 1-6) using outcome-blind simulations based on • causal metrics (95% CI coverage, type 1 error, and power for a causal effect), • and metrics to quantify differences in the medical product development process of each design. • Include a comparison to an RCT, if feasible.

3.4 Discussion

The Roadmap can help investigators to pre-specify a complete analytic design for studies that utilize RWD, choose between study designs, and propose high-quality RWE studies to the FDA and other agencies. We describe the steps of the Roadmap in order to disseminate this methodology to clinical and translational scientists. The case study of Semaglutide and Cardiovascular Outcomes in Chapter 4 demonstrates an application of the Roadmap, explaining specific steps in greater detail.

Past descriptions of the Roadmap have largely been targeted to quantitative scientists[41, 27, 93, 94, 95, 96, 97]. In this Chapter, we focus on intuitive explanations rather than formal mathematical rigor to make these causal inference concepts more accessible to a wide audience. We also emphasize the importance of building a multidisciplinary collaboration, including both clinicians and statisticians, during the study planning phase.

We also introduce an extension of previous versions of the Roadmap to emphasize how outcome-blind simulations may be used not only to compare different statistical estimators but also to evaluate different study designs. This extension aligns with the FDA's Complex Innovative Trial Designs Program guidance for designs that require simulation to estimate type I and II error rates[150], but goes one step further by emphasizing a quantitative comparison to a randomized trial or other feasible RWE designs. The aim of this additional step is to facilitate evaluation of the strengths and weaknesses of each potential approach.

The Roadmap aligns with other recommendations provided in regulatory guidance, as well; these include the FDA's Framework and Draft Guidance documents for RWE that emphasize the quality and appropriateness of the data[59, 151, 152, 153] and the ICH E9(R1) guidance on estimands and sensitivity analysis[107]. The Roadmap is also consistent with other proposed frameworks for RWE generation. Within the field of causal inference, the Roadmap brings together concepts including potential outcomes[28, 111], the careful design of non-experimental studies[102,

103, 105, 106], causal graphs[29, 114, 115, 116, 117] and structural causal models[112], causal identification[29, 112, 154], translation of causal to statistical estimands using the g-formula[29], and methods for estimation and sensitivity analysis[41, 101, 124, 61, 137, 139].

The Roadmap is compatible with many other frameworks, including many that discuss aspects of specific Roadmap steps. Examples include the Target Trial Emulation framework[86, 109], the Patient-Centered Outcomes Research Institute (PCORI) Methodology Standards[88], white papers from the Duke-Margolis Center[87, 155], the REporting of studies Conducted using Observational Routinely-Collected health Data (RECORD) Statement[156], the Structured Preapproval and Postapproval Comparative study design framework[157], and the STaRT-RWE template[89]. The purpose of the Roadmap is not to replace these – and many other – useful sources of guidance, but rather to provide a unified framework that covers the steps necessary to follow a wide range of guidance in a centralized location. Furthermore, while many recommendations for RWE studies list what to think about (e.g., types of biases or considerations for making RWD and trial controls comparable), the Roadmap aims instead to make explicit a process for how to make and report design and analysis decisions that is flexible enough to be applied to any use case along the spectrum from a traditional RCT to a fully observational analysis.

With increasing emphasis by regulatory agencies around the world regarding the importance of RWE[76], the number of studies using RWD that contribute to regulatory decisions is likely to grow over time. Yet a recent review of RWE studies reported that "nearly all [reviewed] studies (95%) had at least one avoidable methodological issue known to incur bias"[158]. By following the Roadmap steps to fully pre-specify an analytic study design, investigators may set themselves up not only to convey relevant information to regulators but also to produce high-quality estimates of causal effects using RWD when possible, with an honest evaluation of whether the proposed study is adequate for making causal inferences.

4 Case Study of Semaglutide and Cardiovascular Outcomes

4.1 Introduction

Semaglutide, a glucagon-like peptide-1 receptor agonist (GLP-1RA) developed as an antihyperglycemic agent, has been shown to improve multiple health outcomes for patients with type 2 diabetes mellitus (T2DM). In the SUSTAIN series of randomized controlled trials (RCTs), injectable semaglutide decreased glycated hemoglobin (HbA1c), body weight, and systolic blood pressure compared to placebo[159], sitagliptin[160], exenatide ER[161], and insulin glargine[162]. In the SUSTAIN 6 trial, injectable semaglutide decreased rates of major adverse cardiovascular events (MACE: defined as death from cardiovascular causes or nonfatal stroke or myocardial infarction (MI)) compared to placebo in patients with high cardiovascular (CV) risk, with an estimated hazard ratio (HR) of 0.74 (95% confidence interval (CI), 0.58 to 0.95)[163]. As a result, the United States Food and Drug Administration (FDA) approved injectable semaglutide for adults with T2DM to improve glycemic control and reduce cardiovascular risk in patients with cardiovascular disease.

Oral semaglutide was subsequently developed and shown, in the PIONEER series of RCTs, to decrease HbA1c compared to placebo[164], empagliflozin[165], and sitagliptin[166], and body weight compared to placebo[164], sitagliptin[166], and liraglutide[167]. To satisfy a pre-approval regulatory requirement for demonstrating cardiovascular safety, the PIONEER 6 RCT was designed to evaluate non-inferiority of oral semaglutide versus placebo[168]. For the primary outcome of MACE, the estimated HR was 0.79 (95% CI, 0.57 - 1.11), a result that was statistically significant for non-inferiority[168]. The evidence obtained through the PIONEER and SUSTAIN trials was not deemed sufficient for FDA approval of oral semaglutide for the secondary indication of cardiovascular risk reduction, prompting initiation of the ongoing SOUL RCT that has enrolled over 9,500 participants[169].

A superiority RCT is a standard choice for evaluating the effect of interest, yet RCTs may also have downsides. For example, clinicians treating placebo arm patients are directed not to prescribe medications of the same class as the active treatment under investigation[169, 170, 171]. Yet in 2019, a joint statement by the American Diabetes Association and the European Association for the Study of Diabetes emphasized that "for patients with type 2 diabetes and established atherosclerotic CV disease ... where MACE is the gravest threat, the level of evidence for MACE benefit is greatest for GLP-1 receptor agonists"[172]. Although none of the trial participants were taking a GLP-1RA at baseline[169], would it not be better for the participants in the placebo arm of SOUL if they were allowed to start a GLP-1RA? This question led us to ask whether alternate trial designs incorporating real-world data (RWD) could decrease the amount of participant-time during which commencement of a GLP1-RA is precluded.

One such design, the hybrid RCT-external data study, has been increasingly utilized for estimating effects of medications for rare diseases[56] and/or pediatric[57] drug approvals when running an adequately powered RCT may not be feasible. Including external data may improve power but may also increase bias. An important subset of hybrid designs, including that considered here[60], estimate the bias that would be introduced by including non-randomized data in the analysis in order to decide whether to estimate the effect of interest based solely on the RCT, based on the pooled RCT and RWD, or based on some weighted combination of the two[60, 173, 3, 11, 12, 13].

In this case study, we evaluate the utility of integrating an external control arm for a common disease, T2DM, with RCT data to estimate benefits for a secondary indication. In this context, an RCT is possible but has potential disadvantages for patient care, while a hybrid RCT-RWD analysis could decrease the amount of time patients are precluded from receiving a GLP1-RA yet raises questions about whether the resulting effect estimate is causal or merely associative. To evaluate these potential tradeoffs, we demonstrate use of the Causal Roadmap[41, 27, 94] – a structured process that helps investigators to pre-specify analytic study designs incorporating RWD – to compare different designs that could be used to estimate the effect of oral semaglutide on cardiovascular outcomes. An overview of the Causal Roadmap may be found in Chapter 3. Specifically, we compare a traditional program of RCTs to a hybrid randomized-RWD study integrating data from the PIONEER 6 non-inferiority trial with external controls from Optum's de-identified Clinformatics® Data Mart Database (CDM) (2007-2022) through simulations that closely mimic these true experiments. We then present results of the real hybrid analysis of PIONEER 6 and CDM for the estimated difference in the risk of a combined outcome of first MI, stroke, or all-cause death with oral semaglutide versus standard-of-care (without a GLP1-RA).

4.2 Methods

Table 4.1: Causal Roadmap Steps for Specification of Study Designs 1-3

Roadmap Step	**Designs 1-2 (RCT Only)**	**Design 3 (RCT + RWD)**
1a. Causal Question/ Causal Estimand	What would the difference in risk of MACE† (defined as death from any cause, nonfatal MI, or nonfatal stroke) within one year be if all patients in a population consistent with the PIONEER 6 inclusion/exclusion criteria and timeframe [168], and with similar healthcare engagement, were prescribed oral semaglutide plus standard-of-care compared to if all patients were prescribed standard-of-care alone, and if censoring had been prevented for all patients? See Appendix B.1 for the mathematical representation of the causal estimand for each study design.	

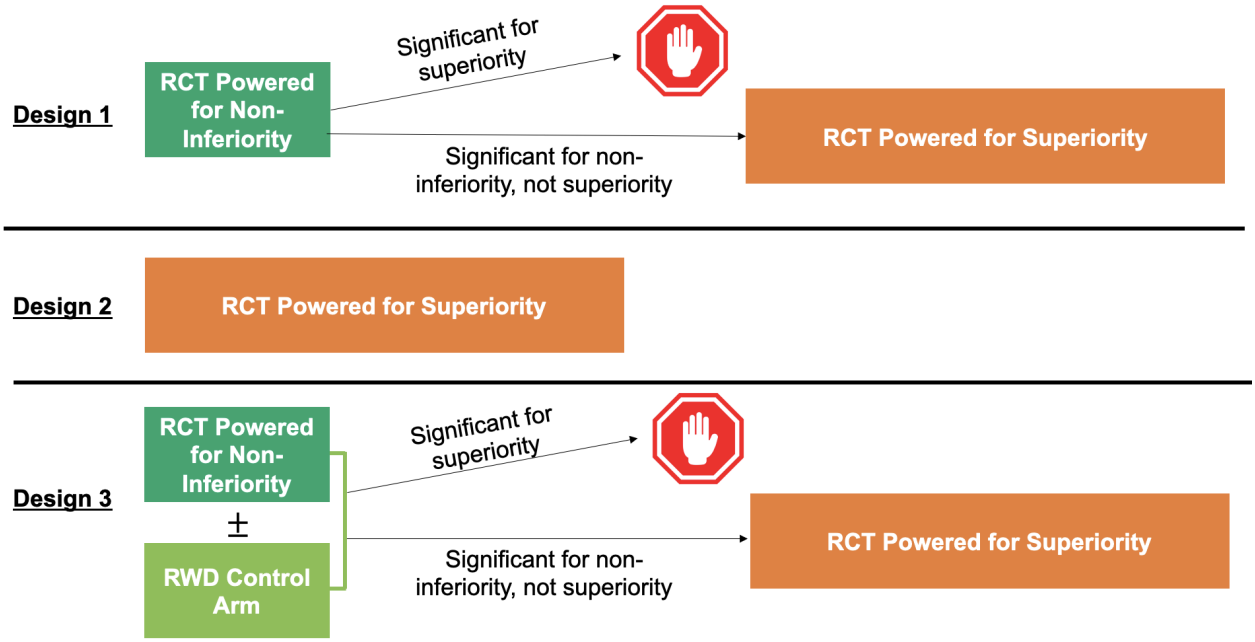
1b. Causal Model	Knowledge about potential shared causes of treatment, censoring, MACE, and participation in the RCT vs RWD, as well as possible causal relations between these variables depicted in Figure 4.2.	
2. Observed Data	Potential data sources: Pioneer 6 RCT, SOUL RCT	Potential data sources: Pioneer 6 RCT, Optum CDM control arm, SOUL RCT
3. Assess Identification	Identification highly likely (non-administrative censoring in PIONEER 6 only 0.3%)	Plausible, though uncertain, that causal gap[††] would be small (see Step 6).
4. Specify Statistical Estimand	Statistical Estimand: Risk difference between treatment and control arms of the trial.	Statistical Estimand: Adjusted risk difference between treatment and control arms, standardized to the covariate distribution in the target population.
5. Statistical Model and Estimator	Statistical Model: Semi-parametric statistical model (incorporating knowledge that treatment was randomized). Estimator: Unadjusted difference in risk between arms.	Statistical Model: Semi-parametric statistical model (incorporating knowledge that treatment in the RCT was randomized). Estimator: Experiment-Selector CV-TMLE
6. Sensitivity Analysis	None given that causal identification assumptions are highly likely to be true.	See Step 6 below.
7. Compare Analytic Designs	See simulation results reported in Step 7 below.	

[†] The revised definition of MACE using all-cause death instead of death from cardiovascular causes was chosen as the primary outcome because cause of death is not available in the RWD.
[††] The causal gap is the difference between the true value of the causal estimand that answers the causal question and the true value of the statistical estimand that we will estimate. [94]

Table 4.1 describes design and analysis plans for three potential study designs for evaluating this question, using the list of Causal Roadmap steps found in Chapter 3; additional detail is provided in the text and appendices. As visually depicted in Figure 4.1, Design 1 is based on what truly

occurred – a non-inferiority trial was run to demonstrate cardiovascular safety of oral semaglutide (PIONEER 6), after which, due to results that were promising but non-significant for superiority, a superiority trial was initiated. Design 2 considers the hypothetical scenario in which only the superiority RCT is run, as might have occurred if superiority had been expected from the start. Design 3 is a hybrid RCT-RWD study in which first a non-inferiority trial potentially augmented with extra RWD controls is run, and the follow-up superiority RCT is only initiated if the hybrid design does not reject the null hypothesis.

Figure 4.1: Diagram of Study Designs 1-3



4.2.1 Step 1a: Define the causal question and estimand

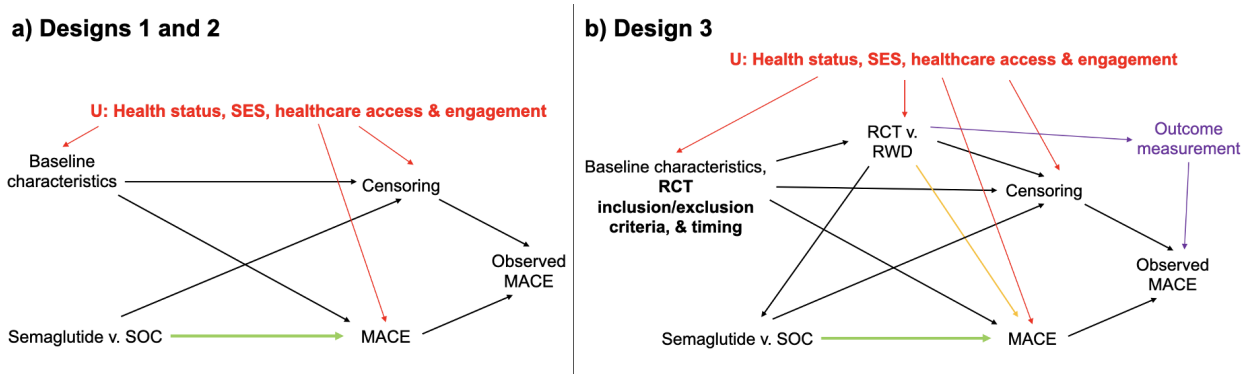
The question for all study designs was: what would the difference in risk of MACE (defined as death from any cause, nonfatal MI, or nonfatal stroke) within one year be if all patients in a population consistent with the PIONEER 6 inclusion/exclusion criteria and timeframe[168], and with similar healthcare engagement, were prescribed oral semaglutide plus standard-of-care compared to if all patients were prescribed standard-of-care alone, and if censoring had been prevented for all patients? The outcome for this case study includes all-cause death (rather than cardiovascular death as in PIONEER 6) because the real-world data do not include cause of death. See Appendix B.1 for the causal estimand.

4.2.2 Step 1b: Specify a causal model

Next, we specify a causal model for each design describing what variables might affect treatment, censoring, or outcomes using the causal graphs[114] shown in Figure 4.2. For the RCTs, only the randomization procedure affects treatment assignment. As depicted in Figure 4.2a, health

status, socioeconomic status, and related issues of healthcare access and engagement, collectively referred to as U, might affect both censoring and MACE. Measured pre-baseline covariates, including age, sex, race, HbA1c, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, estimated glomerular filtration rate (a marker of kidney function), prior MI, prior stroke or transient ischemic attack (TIA), prior heart failure, morbid obesity, and use of glucose-lowering medications, insulin, and CV medications, may account for some aspects of these underlying factors.

Figure 4.2: Causal Graphs for Designs 1-3



SOC: standard-of-care. SES: socioeconomic status. MACE: major adverse cardiovascular events. RWD: real-world data.

In the hybrid Design 3, participation in the RCT versus the real-world system affects treatment because RCT participation is required to receive oral semaglutide if the RWD is concurrent with the pre-approval RCT. Being in the RCT could also modify the effect of treatment or directly affect measured outcomes for reasons including closer monitoring, encouragement of adherence, variation in standard-of-care or placebo effect, or more accurate outcome measurement[1, 174, 175]. The RCT inclusion and exclusion criteria, the timeframe of RCT recruitment, health status, socioeconomic status (SES), and healthcare engagement or access may also affect trial participation.

4.2.3 Step 2: Describe the observed data

Next, we consider the data that are actually observed. Designs 1-2 use data from one or both of the following sources: the PIONEER 6 RCT[168] and the ongoing SOUL RCT[169]. Both trials randomized participants to receive semaglutide or placebo in addition to standard-of-care. The inclusion and exclusion criteria for both trials targeted patients with T2DM and high cardiovascular risk but without unstable disease or recent use of a GLP-1RA, while PIONEER 6 also excluded recent users of pramlintide and dipeptidyl peptidase-4 inhibitors (DPP4i)[169, 171]. Participants were regularly evaluated in person or by phone. Outcomes were adjudicated. Time zero was the time of randomization. The timeframe for the outcome of one year after baseline was selected because no administrative censoring occurred before that time in PIONEER 6.

Finally, the RWD considered in Design 3 comes from Optum's Clinformatics® Data Mart (CDM), which is derived from a database of administrative health claims for members of large commercial and Medicare Advantage health plans. Clinformatics® Data Mart is statistically de-identified under the Expert Determination method consistent with HIPAA and managed according

to Optum® customer data use agreements[176, 177]. CDM administrative claims submitted for payment by providers and pharmacies are verified, adjudicated and de-identified prior to inclusion. This data, including patient-level enrollment information, is derived from claims submitted for all medical and pharmacy health care services with information related to health care costs and resource utilization. The population is geographically diverse, spanning all 50 of the United States.

Consistent with recommendations from the RCT DUPLICATE study, we used naive initiation of a DPP4i (defined as a new prescription following at least 90 days without a previous prescription based on AHFS codes) to enhance comparability of health care access and engagement among RWD compared to RCT controls[178]. Time zero was defined as the first time a participant met the eligibility criteria for PIONEER 6, during a calendar time window contemporaneous with PIONEER 6 recruitment, and was prescribed a DPP4i. Because oral semaglutide was approved following PIONEER 6, we only consider extra RWD control arm participants (although the method we describe can be extended to handle both external control and treatment arms).

Following the RCT Duplicate study[178], we translated and applied as many of the inclusion/exclusion criteria of PIONEER 6 as possible. This involved translating medical histories into ICD-9/ICD-10 diagnosis and procedure codes. To identify GLP1-RA and pramlintide usage, we defined continuous treatment eras as consecutive prescriptions based on AHFS codes with not more than 90-day gaps between them.

For the primary outcome, we identified nonfatal MI/stroke using ICD-9/ICD-10 diagnosis codes for inpatient visits in the first diagnosis position. All-cause death was identified from external sources as provided by Optum. We used fractures as a negative control outcome (discussed below), which we identified using ICD-9/ICD-10 diagnosis codes again for inpatient visits in the first diagnosis position.

Baseline characteristics discussed in Step 1b were determined as follows. Medical history variables were defined as described above. Medication use was identified via claims through AHFS codes, where treatment at baseline corresponded to at least one prescription in the 180 days preceding time zero. Laboratory measurements were identified through LOINC codes, where baseline values were identified as the most recent measurement prior to time zero. If the latest measurement was more than 180 days prior to time zero, then that measurement was deemed missing.

4.2.4 Steps 3-4: Assess identifiability and specify a statistical estimand

We now aim to translate the causal effect of interest (Step 1a) into a statistical estimand – a function of the observed data that we will estimate – based on our knowledge about the processes that generated our data (Step 1b). When using RCT data alone (Designs 1 and 2), we assume no unmeasured common causes of treatment or censoring and MACE, and adequate data support (positivity)[30, 31]. These assumptions are highly likely to hold by design; while it is possible that there are unmeasured common causes of censoring and MACE, because censoring was negligible (0.3% in PIONEER 6), this would be unlikely to impact the results. A design (not considered here) in which we committed to augmenting the RCT data with external control data would require additional assumptions: no unmeasured common causes of selection into the RCT versus the RWD cohort or censoring, and MACE (no U in Figure 4.2b), no effect of RWD vs. RCT participation on MACE other than through effects on treatment assignment (no direct arrow from RCT vs RWD to MACE or outcome measurement in Figure 4.2b), and adequate data support (positivity[30, 31]).

If these assumptions are not close to being satisfied, Design 3 is likely to reject the RWD controls. The following actions were taken to improve the plausibility of these assumptions and there-

fore the likelihood that RWD controls would be integrated in the hybrid design: 1) selecting RWD controls with a similar disease stage and healthcare access and engagement compared to the RCT controls based on those who were prescribed an active comparator medication (DPP4i) and had relevant baseline labs and medical history recorded, 2) restricting RWD controls to the time period of PIONEER 6 recruitment to make standard-of-care more similar, and 3) selecting RWD controls whose baseline characteristics are shared by at least some RCT participants. Note that action 1 restricts the target population for the hybrid analysis to patients with relatively strong access to and engagement with the healthcare system, similar to the types of patients who are able to enroll in an RCT. Note also that use of DPP4i as an active comparator – especially considering that DPP4i use was an exclusion criterion in PIONEER 6 – also requires the assumption, supported by data, that DPP4 inhibitors do not influence cardiovascular outcomes[178]. Action 3 is necessary to avoid a violation of the positivity assumption, but doing so restricts the target population further, preventing generalization beyond the types of patients that were represented in the RCTs.

As always, when observational data are considered, it remains unlikely that the causal identification assumptions are exactly true. Yet with the above considerations, it is now plausible that the bias from integrating RWD will be small. Importantly, however, rather than relying on these assumptions holding, our proposed Design 3 uses an estimator (described in Step 5) in which pre-specified statistical criteria are used to estimate the "causal gap", or difference between the wished-for causal effect and the statistical estimand[94] that may be estimated by pooling RCT and RWD[60]. The estimator only selects the RWD for inclusion when the combined impact of any deviations from these assumptions is unlikely to impact accurate inference[60]. This approach greatly mitigates, but does not eliminate, the threat of misleading inference, which is further evaluated in Steps 6 and 7. Appendix B.1 has mathematical expressions for the statistical estimands that are equivalent to the wished-for causal effects (causal estimands) under the identification assumptions, which are discussed in detail in Appendix B.2.

4.2.5 Step 5: Choose a statistical model and estimator

For all designs, we use a statistical model that avoids any assumptions not firmly grounded in design knowledge (e.g., treatment randomization). In Designs 1 and 2, because censoring was negligible, we estimated the unadjusted risk difference between arms among persons with follow-up through one year. For Design 3, we used the experiment-selector cross-validated targeted maximum likelihood estimator (ES-CVTMLE)[60]. Cross-validated Targeted Maximum Likelihood Estimation (cv-TMLE) is a robust, efficient approach that incorporates machine learning using internal sample splitting (cross-validation)[41, 61, 40, 15]. This allows it to flexibly adjust for covariates without introducing new assumptions, improving precision and potentially reducing bias, while preserving inference[41, 61, 40, 15]. The ES-CVTMLE extends this method to evaluate and integrate external RWD controls[60]. Specifically, it uses pre-specified statistical criteria to evaluate the RWD and integrate them with the RCT data only when their inclusion is unlikely to worsen confidence interval coverage.

The ES-CVTMLE estimates the bias created by augmenting the RCT with RWD in two ways: 1) by comparing the control arms (specifically the conditional mean outcomes) between the two settings and 2) through the use of a negative control outcome (NCO)[60]. From the available options, we chose fractures as an NCO because it is generally serious enough to require medical attention for those with access, is associated with SES[179], and is recorded in a manner similar to the primary outcome. Studies prospectively designed to include an NCO could consider alternate choices.

In this case study, we chose the ES-CVTMLE as the estimator for the hybrid design because it relies on few statistical assumptions, incorporates an estimate of bias based on an NCO, and adjusts confidence interval widths based on the estimated magnitude of bias. While the focus of this case study is on using simulations to compare study designs, similar simulations could also be used to compare different potential estimators. Appendices B.3-6 provide further details about estimation.

4.2.6 Step 6: Specify a procedure for sensitivity analysis

For the RCT-only analyses (Designs 1 and 2), because identification assumptions are likely to hold and we avoid any unsupported statistical assumptions, there are few threats to inference, and sensitivity analyses may be unnecessary. In Design 3, for the hybrid RCT-RWD analysis, the ES-CVTMLE is designed to conservatively protect inference without additional assumptions. However, as with all estimators that aim to estimate bias from including external data in order to decide whether or how to include RWD in a hybrid analysis[173, 3, 11, 12, 13, 10], it still runs a small but real risk of failing to reject inappropriate external controls. To evaluate the extent to which this threatens inference, a tipping point analysis can be carried out to evaluate how much of a causal gap for the combined RCT-RWD study is needed to make the confidence interval from Step 5 non-significant[180, 140, 124, 125]. Additional simulation-based approaches, incorporating estimates of the plausible causal gap, can also be used to augment sensitivity analyses.

4.2.7 Step 7: Compare analytic designs using simulations

We compare Designs 1-3 using simulations that closely mimic our true study designs (Appendix B.4). For Designs 1 and 3, we simulate data from a small RCT aiming to mimic PIONEER 6 (RCT1). For Design 1, we use an unadjusted estimate of the difference in risk between arms of RCT1. For Design 3, we consider not only the simulated RCT1 but also a simulated "real-world" dataset aiming to mimic CDM, and we use the ES-CVTMLE to estimate the risk difference. In both Designs 1 and 3, if the null hypothesis is rejected at this first stage, then we use this initial estimate as our final effect estimate. If the null hypothesis is not rejected, then we simulate data from a larger trial aiming to mimic a superiority trial (RCT2) and estimate the risk difference using an unadjusted estimate. In Design 2, we simply simulate RCT2 and use the unadjusted effect estimate.

Because we consider non-randomized data in Design 3, it is possible that the confidence interval coverage will be lower than in Designs 1-2 if the causal identification assumptions from Step 3 are violated. Yet it may be acceptable if, in some contexts (i.e., bias of a certain magnitude), coverage is below 95% if there is sufficient benefit to patients of Design 3 over Designs 1-2. While there are many potential ways to quantify benefit to patients, for this case study, we estimate the number of patient-years during which participants are precluded from starting a GLP1-RA (by being in an RCT control arm), averaged over 1000 iterations of this simulation. We also report power to reject the null hypothesis using $\alpha = 0.05$.

We evaluate Design 3 when the magnitude of bias introduced by including the simulated RWD in the analysis is zero and when it is one of ten potential magnitudes in either the positive or negative direction, ranging up to $\pm 2.1\%$. Appendix B.4 describes the rationale for this maximum bias. In this primary simulation, the effect of unmeasured factors causing bias is the same on the relationship between semaglutide and MACE as it is on the relationship between semaglutide and the NCO. Appendix B.5 shows the results of the same simulation both when the NCO is not considered (leading to more conservative inference) and when the NCO is considered but the unmeasured factors causing bias for the relationship between semaglutide and MACE have no effect on the NCO,

mimicking a worst-case scenario for violations of both the causal identification assumptions and the assumptions needed to measure bias using an NCO.

4.3 Results

4.3.1 Simulation results

Figure 4.3 shows the results of 1000 iterations of the simulation comparing Designs 1-3. Design 1 and Design 2 had similar characteristics. Simulated Design 1 had 95% CI coverage of 0.949, power of 0.842, and an average of 4,765 patient-years during which a GLP1-RA was precluded. Simulated Design 2 had coverage of 0.955, power of 0.764, and an average of 4,750 patient-years during which a GLP1-RA was precluded.

Figure 4.3: Simulation Results by Study Design with Different Amounts of RWD Bias

Caption: Purple represents 10 simulated magnitudes of bias away from the null. Pink represents 10 simulated magnitudes of bias towards the null.

The tradeoffs between Design 3 and Designs 1 and 2 depended on the direction of bias introduced by the RWD. With unbiased simulated RWD, Design 3 had coverage of 0.942, had power of 0.866, and resulted in an average of 394 fewer participant-years during which patients were precluded from starting a GLP1-RA compared to Design 1. In other words, on average, 8.3% fewer people would have spent one year during which their doctor avoided prescribing a GLP1-RA if Design 3 were chosen and unbiased RWD were available compared to if Design 1 were chosen.

Positive bias (towards the null) represents scenarios in which the introduction of RWD lowers the estimated risk of MACE among control arm participants. This could happen if MACE was not well-recorded in the RWD. Simulated positive bias led to coverage ranging from 0.945 to 0.948, power ranging from 0.842 to 0.847, and an average of 0 to 114 extra participant-years during which prescription of a GLP1-RA was discouraged compared to Design 1, with the largest increases for intermediate magnitudes of bias. The increase in person-years without GLP1-RA access occurred because RWD with bias towards the null were included in a small number of simulation iterations, triggering a second RCT. A study comparing outcomes recorded in CDM to outcomes recorded following an RCT protocol could be conducted to assess the likelihood that the proposed hybrid design would truly result in this spectrum of positive RWD bias.

Negative bias (away from the null) represents scenarios in which the introduction of RWD raises the estimated risk of MACE among control arm participants. This could happen if participants in the real world had worse health outcomes than trial participants due to differences described in Section 2 and these differences were not adequately detected through comparison of RWD and RCT control arms or via negative control. Simulated negative bias led to coverage ranging from 0.927 to 0.948, power ranging from 0.847 to 0.873, and an average of 0 to 542 fewer participant-years during which prescription of a GLP1-RA was discouraged compared to Design 1. For comparison, if one had used a naive estimator that simply assumed that external control data were appropriate to integrate (i.e., the stringent assumptions discussed in Step 3 held), and simply pooled the RCT and most biased real-world data, coverage would have been 0.126. The ES-CVTMLE thus provided significant (though imperfect) protection against integration of biased RWD in this simulation.

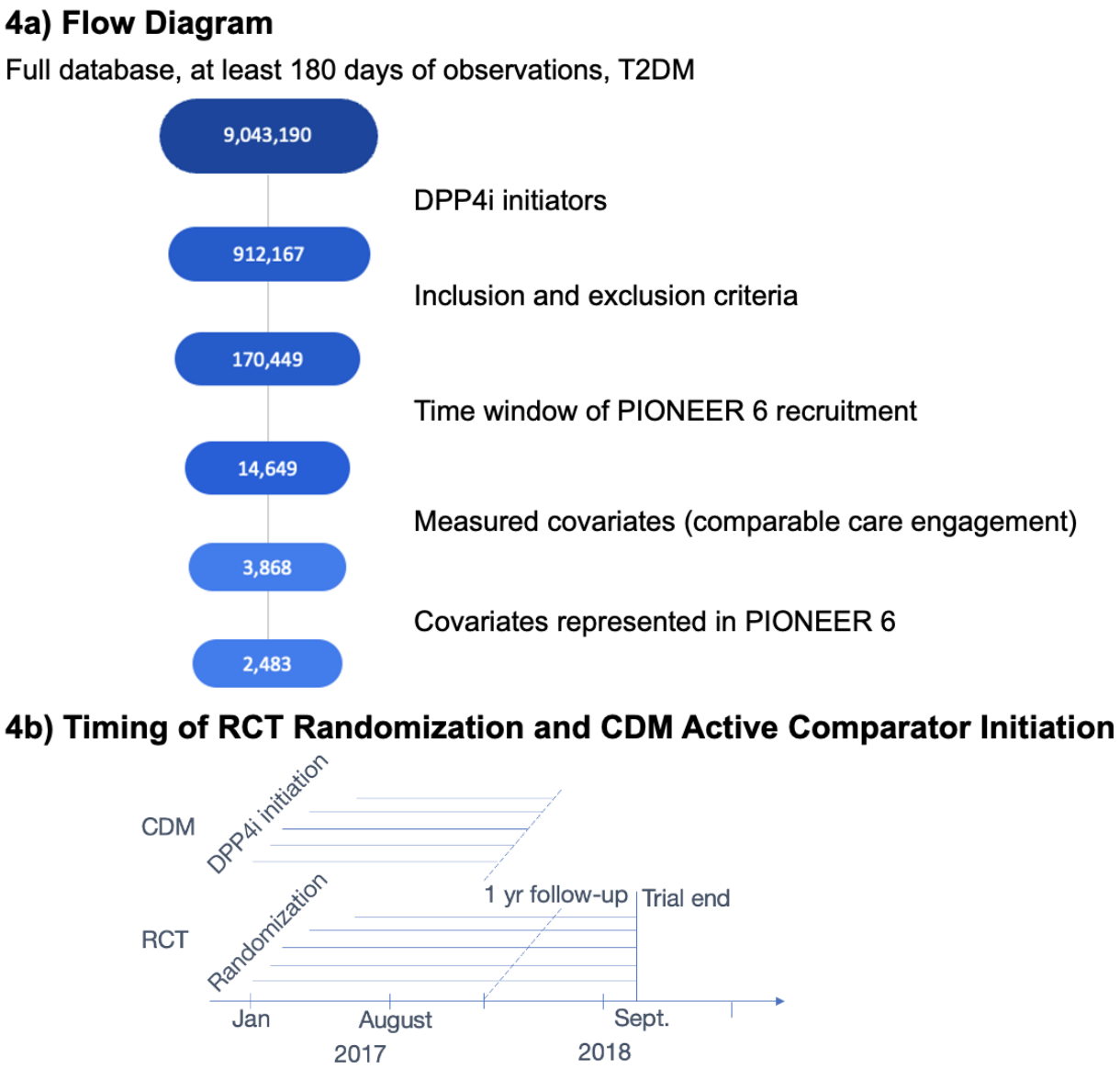
The possibility of bias away from the null is plausible, though quantifying how much bias might be expected within the range represented by this simulation (0-2.1$\%$) is challenging. Nonetheless, by objectively quantifying these differences between proposed designs, investigators can explicitly discuss these tradeoffs with stakeholders such as patient groups and regulatory agencies when deciding which trial design to choose.

4.3.2 Real data analysis: The estimated effect of oral semaglutide on MACE from PIONEER 6, considering augmentation with additional CDM RWD controls

The actual results of Designs 1 and 2 await completion of the SOUL trial. Below, we carry out Design 3 using data from PIONEER 6 and the CDM external control arm described in Step 2 above. We also report the results of an unadjusted estimator for the difference in the risk of MACE among PIONEER 6 active and control arm participants.

After applying the inclusion and exclusion criteria described in Step 2 and depicted in Figure 4.4, the CDM cohort consisted of 2483 participants. Table 4.2 lists baseline demographics, medical history, medication use, outcome missingness, and MACE and NCO event rates for the PIONEER 6 semaglutide and placebo arms, as well as for the CDM external control arm. The outcome was missing for 0.3% of PIONEER 6 participants and 16% of CDM participants. The negative control outcome was missing for 0.3% of PIONEER6 participants and 17% of CDM participants. Compared to the PIONEER 6 control arm, the CDM controls were slightly older (more participants in the 70-80 year-old range), had a higher percentage of females, had a lower proportion of previous MI or stroke but a higher proportion of previous heart failure, and had a different distribution of baseline medication use.

Figure 4.4: Selection of CDM External Control Group



Caption: T2DM: Type 2 Diabetes Mellitus. DPP4i: Dipeptidyl Peptidase 4 inhibitor.

Table 4.2: Baseline Characteristics, Outcome Missingness, and Event Rates for PIONEER 6 and CDM

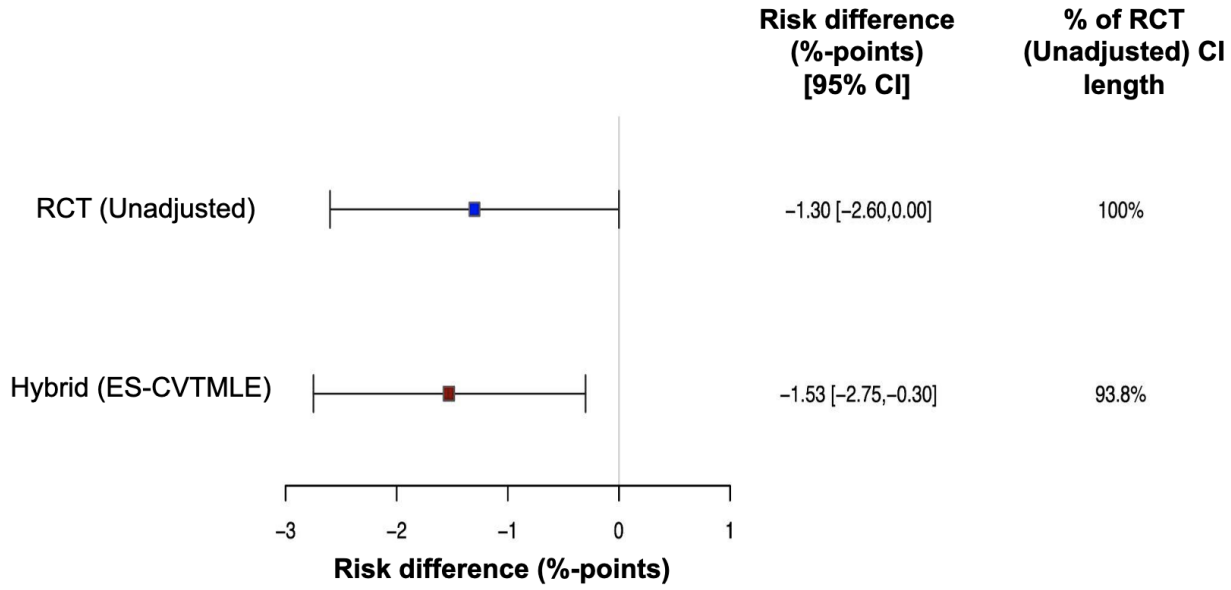
	CDM RWD control arm (n=2483)	PIONEER 6 placebo arm (n=1564)	PIONEER 6 semaglutide arm (n=1574)
MACE rate - %	4.5	4.2	2.9
NCO Rate - %	0.73	0.77	0.45
Age - years, mean (SD)	69.2 (6.3)	66.4 (7.1)	65.9 (7.1)
Female sex - %	42.7	31.2	31.9
Race			
White - %	44	72	72
Black - %	11	7	6
Other - %	45	21	22
HbA1c - %, mean (SD)	8.0 (1.4)	8.2 (1.6)	8.2 (1.6)
LDL cholesterol - mg/dl, mean (SD)	84.1 (28.3)	84.8 (32.4)	83.9 (34.0)
HDL cholesterol - mg/dl, mean (SD)	44.1 (9.6)	41.6 (10.7)	41.9 (11.0)
eGFR - ml/min/1.73 m2, mean (SD)	74.3 (19.4)	74.2 (20.9)	74.2 (21.1)
Previous MI - %	13.9	36.9	35.3
Previous stroke/TIA - %	11.8	16.6	15.2
Previous heart failure - %	20.4	12.4	11.9
Morbid obesity - %	16.4	12.3	12.2
Glucose-lowering medication (metformin, SU, TZD, SGLT2i)	73.0	83.9	83.9
Insulin	14.9	61.2	61.2
Cardiovascular medication (antihypertensives, lipid-lowering, anti-thrombosis, diuretics)	91.5	98.9	98.9
Outcome missingness - %	15.8	0.3	0.3

NCO missingness -%	17.2	0.3	0.3

LDL: low-density lipoprotein, HDL: high-density lipoprotein, eGFR: estimated glomerular filtration rate, MI: myocardial infarction, TIA: transient ischemic attack, SU: sulfonylurea, TZD: thiazolidinedione, SGLT2i: sodium/glucose cotransporter-2 inhibitor

As shown in Figure 4.5, the estimated difference in the risk of MACE by 1 year based on the unadjusted estimator conducted using PIONEER 6 data alone was -1.30%-points (95% CI -2.60 to 0.00%-points). This result is closer to statistical significance than the primary result reported for the PIONEER 6 trial (hazard ratio 0.79; 95% CI 0.57 to 1.11[168]) because the primary analysis a) evaluated the hazard ratio including all timepoints instead of the risk difference by one year and b) evaluated a composite outcome that included death from cardiovascular causes instead of death from all causes. Nonetheless, the confidence interval for the result of this modified analysis still includes zero.

Figure 4.5: Estimated Difference in 1-Year Risk of MACE for PIONEER 6 and Hybrid Design



Caption: CI: Confidence Interval. ES-CVTMLE: Experiment-Selector Cross-Validated Targeted Maximum Likelihood Estimator.

Hybrid Design 3 resulted in an estimated risk difference of -1.53%-points (95% CI -2.75 to -0.30%-points), providing evidence in support of the superiority of oral semaglutide versus standard-of-care for the prevention of MACE. The two primary differences between the risk difference estimates from PIONEER 6 alone compared to the hybrid analysis are narrower confidence intervals and a small negative shift in the point estimate. Narrower confidence intervals are expected as the CDM RWD were included in the analysis in 84% of internal sample splits, leading to increased efficiency.

The small shift in the point estimate could occur for three main reasons. First, the magnitude of the shift is well within what might be expected by chance alone. Second, the shift may be due to subtle changes in the target population that arise from including external controls (even those with equivalent eligibility criteria (Table 4.2)). Finally, as discussed above and illustrated using simulations, the potential for some residual bias remains.

Simulations, such as those presented in Step 7, can help stakeholders to weigh the tradeoffs between the benefits and downsides to patients of the different study designs against the risk of misleading inference. Had this design been proposed prior to running SOUL, explicit evaluation of these tradeoffs could have facilitated discussions of whether the evidence produced by the proposed hybrid design would be sufficient to inform decisions regarding label extensions for oral semaglutide for the secondary indication of cardiovascular risk reduction.

4.4 Discussion

In this case study of the effect of oral semaglutide on major adverse cardiovascular events, we demonstrate an application of the Causal Roadmap to a hybrid RCT-RWD trial that considers integration of data from multiple sources. We also implement the extension proposed in Chapter 3 to use the Causal Roadmap to compare different potential study designs using simulations. Both the FDA guidance on complex innovative trial designs[150] and the FDA guidance on adaptive designs[135] suggest the utility of simulations for comparing alternative design choices. The simulated results demonstrate that selection of one potential design over another may depend on the direction of the bias introduced by RWD and suggest the importance of conducting studies to evaluate the potential for bias with integration of different RCT and RWD sources.

Our estimate of the difference in the risk of MACE by one year with oral semaglutide versus standard-of-care attained statistical significance, providing evidence in support of the superiority of oral semaglutide as compared to standard-of-care alone. Regulatory decisions regarding whether to extend the label of oral semaglutide to include the secondary indication of cardiovascular risk reduction will await the results of the SOUL trial. Nonetheless, hybrid RCT-RWD studies could potentially be used in the future to provide additional relevant information to regulatory agencies for secondary indications for a wide variety of disease processes – in this case, a common adult disease rather than a pediatric or rare disease as have been highlighted in previous applications of hybrid trial methodologies[56, 57].

While this study aimed to quantify tradeoffs between three proposed designs, other approaches could be considered. Hybrid RCT-RWD designs may adapt the probability of randomization to active treatment based on the efficiency gains that are achieved by integrating RWD and also on the probability of superiority compared to placebo[3, 147, 181], potentially leading to even less patient-time on an inferior product. Power could also have been higher if oral semaglutide had been available in the CDM dataset for the specified time period – resulting in integration of both extra treatment and control arm participants – or if more RWD controls were available.

A key limitation of this study is that it was planned after PIONEER 6, and thus the simulations and analysis plan were not pre-specified without knowledge of trial results. In RWD studies aiming to support policy and/or regulatory decision-making, all design and analysis decisions should be pre-specified before effects are estimated from any of the data sources that are considered. The Causal Roadmap supports a rigorous design process and reporting structure to ensure pre-specification of components needed to support the validity of causal inferences drawn from

such designs. The current case study provides a detailed worked example of this process.

This study reports one of many potential metrics aimed at quantifying the benefits and drawbacks of different study designs from the perspective of patients. While recommendations have been proposed regarding the elicitation of patient perspectives to inform medical product development [182, 149], further guidance on the most relevant metrics of patient benefit as well as best practices for collaboratively weighing tradeoffs between different metrics of design performance is warranted. Decisions regarding these tradeoffs are context-specific[150], but following the Causal Roadmap may help investigators to collate and quantify the information that is most relevant for discussions with stakeholders. For this case study of semaglutide and cardiovascular outcomes, application of the Causal Roadmap prompts us to ask whether more patients could have benefited from receiving a GLP1-RA sooner if a hybrid RCT-RWD approach had been taken.

5 Bibliography

[1] Pocock S. J. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases.* 1976;29(3):175–188.

[2] Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics.* 2014;13(1):41–54.

[3] Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information: Robust Meta-Analytic-Predictive Priors. *Biometrics.* 2014;70(4):1023–1032.

[4] Jahanshahi M, Gregg K, Davis G, et al. The Use of External Controls in FDA Regulatory Decision Making. *Therapeutic Innovation & Regulatory Science.* 2021;55(5):1019-1035.

[5] Dejardin D, Delmar P, Warne C, Patel K, Rosmalen J, Lesaffre E. Use of a historical control group in a noninferiority trial assessing a new antibacterial treatment: A case study and discussion of practical implementation aspects. *Pharmaceutical Statistics.* 2018;17(2):169–181.

[6] Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences.* 2016;113(27):7345–7352.

[7] Ibrahim J, Chen M.-H. Power Prior Distributions for Regression Models. *Statistical Science.* 2000;15(1):46–60.

[8] Hobbs B. P, Sargent D. J, Carlin B. P. Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. *Bayesian Analysis.* 2012;7(3).

[9] Rosenman E. T, Basse G, Owen A. B, Baiocchi M. Combining observational and experimental datasets using shrinkage estimators. *Biometrics.* 2023:biom.13827.

[10] Yang S, Gao C, Zeng D, Wang X. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 2023:qkad017.

[11] Chen S, Zhang B, Ye T. Minimax Rates and Adaptivity in Combining Experimental and Observational Data. *arXiv.* Preprint posted online September 22, 2021. arXiv:2109.10522.

[12] Cheng D, Cai T. Adaptive Combination of Randomized and Observational Data. *arXiv.* Preprint posted online November 29, 2021. arXiv:2111.15012.

[13] Oberst M, D'Amour A, Chen M, Wang Y, Sontag D, Yadlowsky S. Bias-robust Integration of Observational and Experimental Estimators. *arXiv.* Preprint posted online May 21, 2022. arXiv:2205.10467.

[14] Zheng W, van der Laan M. Asymptotic Theory for Cross-Validated Targeted Maximum Likelihood Estimation. Working Paper 273. 2010. University of California, Berkeley. `https://biostats.bepress.com/ucbbiostat/paper273/`.

[15] Hubbard A. E, Kherad-Pajouh S, Laan M. J. Statistical Inference for Data Adaptive Target Parameters. *The International Journal of Biostatistics.* 2016;12(1):3–19.

[16] Sofer T, Richardson D. B, Colicino E, Schwartz J, Tchetgen Tchetgen E. J. On Negative Outcome Control of Unobserved Confounding as a Generalization of Difference-in-Differences. *Statistical Science.* 2016;31(3):348-361.

[17] Shi X, Miao W, Tchetgen E. T. A Selective Review of Negative Control Methods in Epidemiology. *Current Epidemiology Reports.* 2020;7(4):190–202.

[18] Galwey N. W. Supplementation of a clinical trial by historical control data: is the prospect of dynamic borrowing an illusion?. *Statistics in Medicine.* 2017;36(6):899–916.

[19] Cuffe R. L. The inclusion of historical control data may reduce the power of a confirmatory study. *Statistics in Medicine.* 2011;30(12):1329–1338.

[20] Harun N, Liu C, Kim M. Critical appraisal of Bayesian dynamic borrowing from an imperfectly commensurate historical control. *Pharmaceutical Statistics.* 2020;19(5):613–625.

[21] Li W, Liu F, Snavely D. Revisit of test-then-pool methods and some practical considerations. *Pharmaceutical Statistics.* 2020;19(5):498–517.

[22] Green E, Strawderman W. A James-Stein Type Estimator for Combining Unbiased and Possibly Biased Estimators. *Journal of the American Statistical Association.* 1991;86(416):1001–1006.

[23] Stein C. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* 1956;3.1:197–206.

[24] Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology.* 2010;21(3):383–388.

[25] Shi X, Miao W, Nelson J. C, Tchetgen Tchetgen E. J. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2020;82(2):521–540.

[26] Miao W, Shi X, Tchetgen E. T. A Confounding Bridge Approach for Double Negative Control Inference on Causal Effects. *arXiv.* Preprint posted online September 18, 2020. arXiv:1808.04945.

[27] Petersen M. L, van der Laan M. J. Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology.* 2014;25(3):418–426.

[28] Neyman J. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. English translation by D.M. Dabrowska and T.P. Speed (1990). *Statistical Science.* 1923;5:465–480.

[29] Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling.* 1986;7(9-12):1393–1512.

[30] Hernan M. A. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health.* 2006;60(7):578–586.

[31] Petersen M. L, Porter K. E, Gruber S, Wang Y, van der Laan M. J. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research.* 2012;21(1):31–54.

[32] Rudolph K. E, van der Laan M. J. Robust Estimation of Encouragement Design Intervention Effects Transported Across Sites. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 2017;79(5):1509–1525.

[33] Dahabreh I. J, Robertson S. E, Tchetgen E. J, Stuart E. A, Hernán M. A. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics.* 2019;75(2):685–694.

[34] Dahabreh I. J, Haneuse S. J.-P. A, Robins J. M, et al. Study designs for extending causal inferences from a randomized trial to a target population. *American Journal of Epidemiology.* 2021;190(8):1632-1642.

[35] Hartman E, Grieve R, Ramsahai R, Sekhon J. S. From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society Series A.* 2015;178(3):757–778.

[36] Balzer L. B. "All Generalizations Are Dangerous, Even This One."—Alexandre Dumas:. *Epidemiology.* 2017;28(4):562–566.

[37] Dahabreh I. J, Robins J. M, Haneuse S. J.-P. A, Hernán M. A. Generalizing causal inferences from randomized trials: Counterfactual and graphical identification. *arXiv.* Preprint posted online June 26, 2019. arXiv:1906.10792.

[38] Balzer L, Petersen M. L, van der Laan M. Targeted Estimation and Inference for the Sample Average Treatment Effect in Trials with and without Pair-Matching. *Statistics in Medicine.* 2016;35(21):3717-3732.

[39] van der Laan M. J, Luedtke A. R. Targeted Learning of the Mean Outcome under an Optimal Dynamic Treatment Rule. *Journal of Causal Inference.* 2015;3(1):61–95.

[40] van der Laan M. J, Rubin D. Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics.* 2006;2(1).

[41] van der Laan M. J, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer. 2011.

[42] Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable. *Statistical Science.* 2007;22(4):544-559.

[43] Rotnitzky A, Lei Q, Sued M, Robins J. M. Improved double-robust estimation in missing data and causal inference models. *Biometrika.* 2012;99(2):439–456.

[44] Tran L, Yiannoutsos C, Wools-Kaloustian K, Siika A, van der Laan M, Petersen M. Double Robust Efficient Estimators of Longitudinal Treatment Effects: Comparative Performance in Simulations and a Case Study. *The International Journal of Biostatistics.* 2019;15(2):20170054.

[45] Gruber S, van der Laan M. tmle: An R Package for Targeted Maximum Likelihood Estimation. *Journal of Statistical Software.* 2012;51(13):1–35.

[46] Weber S, Li Y, Seaman J. W, Kakizume T, Schmidli H. Applying Meta-Analytic-Predictive Priors with the R Bayesian Evidence Synthesis Tools. *Journal of Statistical Software.* 2021;100(19):1–32.

[47] Marso S. P, Daniels G. H, Brown-Frandsen K, et al. Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes. *New England Journal of Medicine.* 2016;375(4):311–322.

[48] Zinman B, Nauck M. A, Bosch-Traberg H, et al. Liraglutide and Glycaemic Outcomes in the LEADER Trial. *Diabetes Therapy.* 2018;9(6):2383–2392.

[49] Blasco-Blasco M, Puig-García M, Piay N, Lumbreras B, Hernández-Aguado I, Parker L. A. Barriers and facilitators to successful management of type 2 diabetes mellitus in Latin America and the Caribbean: A systematic review. *PLoS ONE.* 2020;15(9):e0237542.

[50] Avilés-Santa M. L, Monroig-Rivera A, Soto-Soto A, Lindberg N. M. Current State of Diabetes Mellitus Prevalence, Awareness, Treatment, and Control in Latin America: Challenges and Innovative Solutions to Improve Health Outcomes Across the Continent. *Current Diabetes Reports.* 2020;20(11):62.

[51] Venkitachalam L, Wang K, Porath A, et al. Global Variation in the Prevalence of Elevated Cholesterol in Outpatients With Established Vascular Disease or 3 Cardiovascular Risk Factors According to National Indices of Economic Development and Health System Performance. *Circulation.* 2012;125(15):1858–1869.

[52] Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software.* 2011;45(3):1–67.

[53] Enea M. *speedglm: Fitting Linear and Generalized Linear Models to Large Data Sets.* 2022. R package version 0.3-4.

[54] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software.* 2010;33(1):1-22.

[55] Milborrow S. *earth: Multivariate Adaptive Regression Splines.* 2021. R package version 5.3.1. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper.

[56] U.S. Food and Drug Administration . CID Case Study: External Control in Diffuse B-Cell Lymphoma. 2022. `https://www.fda.gov/media/155405/download`.

[57] Brunner H. I, Abud-Mendoza C, Viola D. O, et al. Safety and efficacy of intravenous belimumab in children with systemic lupus erythematosus: results from a randomised, placebo-controlled trial. *Annals of Rheumatic Diseases.* 2020;79(10):1340–1348.

[58] Bunnik E. M, Aarts N. What do patients with unmet medical needs want? A qualitative study of patients' views and experiences with expanded access to unapproved, investigational treatments in the Netherlands. *BMC Medical Ethics.* 2019;20(1):80.

[59] U.S. Food and Drug Administration . Framework for FDA's Real-World Evidence Program. 2018. `https://www.fda.gov/media/120060/download`.

[60] Dang L. E, Tarp J. M, Abrahamsen T. J, et al. A Cross-Validated Targeted Maximum Likelihood Estimator for Data-Adaptive Experiment Selection Applied to the Augmentation of RCT Control Arms with External Data. *arXiv.* Preprint posted online October 11, 2022. arXiv:2210.05802.

[61] van der Laan M. J, Polley E. C, Hubbard A. E. Super Learner. *Statistical Applications in Genetics and Molecular Biology.* 2007;6(1).

[62] Hubbard A, van der Laan M. Mining with inference: Data adaptive target parameters. In: Buhlmann P, Drineas P, Kane M, van der Laan M. , eds. *Handbook of Big Data.* Boca Raton, FL: Chapman & Hall/CRC. 2016. 455-468.

[63] Lendle S. D, Schwab J, Petersen M. L, van der Laan M. J. ltmle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data. *Journal of Statistical Software.* 2017;81(1):1–21.

[64] van der Laan M, Coyle J. R, Hejazi N. S, Malenica I, Phillips R, Hubbard A. Targeted Learning in R: Causal Data Science with the tlverse Software Ecosystem. 2022. `https://tlverse.org/tlverse-handbook/`.

[65] Luby S. P, Rahman M, Arnold B. F, et al. Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised controlled trial. *The Lancet Global Health.* 2018;6(3):e302–e315.

[66] Polley E, LeDell E, Kennedy C, van der Laan M. *SuperLearner: Super Learner Prediction.* 2021. R package version 2.0-28.

[67] Phillips R. V, van der Laan M. J, Lee H, Gruber S. Practical considerations for specifying a super learner. *International Journal of Epidemiology.* 2023;00(00):1-10.

[68] Coyle J, Hejazi N, Malenica I, Phillips R. *origami: Generalized Framework for Cross-Validation.* 2022. R package version 1.0.7.

[69] Gruber S, van der Laan M. J. A Targeted Maximum Likelihood Estimator of a Causal Effect on a Bounded Continuous Outcome. *The International Journal of Biostatistics.* 2010;6(1).

[70] Gruber S, Phillips R. V, Lee H, van der Laan M. J. Data-Adaptive Selection of the Propensity Score Truncation Level for Inverse-Probability–Weighted and Targeted Maximum Likelihood Estimators of Marginal Point Treatment Effects. *American Journal of Epidemiology.* 2022;191(9):1640–1651.

[71] Arnold B. F, Null C, Luby S. P, Colford J. M. Implications of WASH Benefits trials for water and sanitation – Authors' reply. *The Lancet Global Health.* 2018;6(6):e616–e617.

[72] World Bank . Bangladesh Poverty Assessment: Assessing a Decade of Progress in Reducing Poverty 2000-2010. 2013. Dhaka, Bangladesh. `https://documents1.worldbank.org/curated/en/109051468203350011/pdf/Bangladesh-Poverty-assessment-assessing-a-decade-of-progress-in-reducing-poverty-2000-2010.pdf`.

[73] Rudrapatna V. A, Butte A. J. Opportunities and challenges in using real-world data for health care. *Journal of Clinical Investigation.* 2020;130(2):565–574.

[74] 21st Century Cures Act, H.R. 34, 114th Congress. 2016.

[75] National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Sciences Policy; Forum on Drug Discovery, Development, and Translation . Perspectives on Real-World Evidence. In: Shore C, Gee A. W, Kahn B, Forstag E. H. , eds. *Examining the Impact of Real-World Evidence on Medical Product Development: Proceedings of a Workshop Series.* Washington, DC: National Academies Press (US). 2019.

[76] Burns L, Roux N. L, Kalesnik-Orszulak R, et al. Real-World Evidence for Regulatory Decision-Making: Guidance From Around the World. *Clinical Therapeutics.* 2022;44(3):420–437.

[77] Baumfeld Andre E, Reynolds R, Caubel P, Azoulay L, Dreyer N. A. Trial designs using real-world data: The changing landscape of the regulatory approval process. *Pharmacoepidemiology and Drug Safety.* 2020;29(10):1201–1212.

[78] U.S. Food and Drug Administration . Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products: Guidance for Industry. 2023. `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products`.

[79] Lavertu A, Vora B, Giacomini K. M, Altman R, Rensi S. A New Era in Pharmacovigilance: Toward Real-World Data and Digital Monitoring. *Clinical Pharmacology and Therapeutics.* 2021;109(5):1197–1202.

[80] Ringel M. S, Scannell J. W, Baedeker M, Schulze U. Breaking Eroom's Law. *Nature Reviews Drug Discovery.* 2020;19(12):833–834.

[81] U.S. Food and Drug Administration . Antibacterial Therapies for Patients with an Unmet Medical Need for the Treatment of Serious Bacterial Diseases: Guidance for Industry. 2017. `https://www.fda.gov/media/86250/download`.

[82] Barda N, Dagan N, Cohen C, et al. Effectiveness of a third dose of the BNT162b2 mRNA COVID-19 vaccine for preventing severe outcomes in Israel: an observational study. *The Lancet.* 2021;398(10316):2093–2100.

[83] Monge S, Rojas-Benedicto A, Olmedo C, et al. Effectiveness of mRNA vaccine boosters against infection with the SARS-CoV-2 omicron (B.1.1.529) variant in Spain: a nationwide cohort study. *The Lancet Infectious Diseases.* 2022;22(9):1313–1320.

[84] Dickerman B. A, Gerlovin H, Madenci A. L, et al. Comparative Effectiveness of BNT162b2 and mRNA-1273 Vaccines in U.S. Veterans. *New England Journal of Medicine.* 2022;386(2):105–115.

[85] Dagan N, Barda N, Biron-Shental T, et al. Effectiveness of the BNT162b2 mRNA COVID-19 vaccine in pregnancy. *Nature Medicine.* 2021;27(10):1693–1695.

[86] Hernán M. A, Robins J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology.* 2016;183(8):758–764.

[87] Berger M, Overhage M, Daniel G, et al. A Framework for Regulatory Use of Real-World Evidence. 2017. Duke-Margolis Center for Health Policy. `https://healthpolicy.duke.edu/sites/default/files/2020-08/rwe_white_paper_2017.09.06.pdf`.

[88] Patient-Centered Outcomes Research Institute . PCORI Methodology Standards. 2019. `https://www.pcori.org/research/about-our-research/research-methodology/pcori-methodology-standards`.

[89] Wang S. V, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ.* 2021:m4856.

[90] Arlett P, Kjær J, Broich K, Cooke E. Real-World Evidence in EU Medicines Regulation: Enabling Use and Establishing Value. *Clinical Pharmacology and Therapeutics.* 2022;111(1):21–23.

[91] National Institute for Health and Care Excellence . NICE real-world evidence framework. 2022. `www.nice.org.uk/corporate/ecd9`.

[92] Elm E, Altman D. G, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Epidemiology.* 2007;18(6):800–804.

[93] Ho M, van der Laan M, Lee H, et al. The Current Landscape in Biostatistics of Real-World Data and Evidence: Causal Inference Frameworks for Study Design and Analysis. *Statistics in Biopharmaceutical Research.* 2021;15(1):43-56.

[94] Gruber S, Phillips R. V, Lee H, Ho M, Concato J, Laan M. J. Targeted learning: Towards a future informed by real-world evidence. *Statistics in Biopharmaceutical Research.* 2023;00(00):1-15.

[95] Petersen M. L. Commentary: Applying a Causal Road Map in Settings with Time-dependent Confounding. *Epidemiology.* 2014;25(6):898–901.

[96] Balzer L, Petersen M, van der Laan M. Tutorial for Causal Inference. In: Buhlmann P, Drineas P, Kane M, van der Laan M. , eds. *Handbook of Big Data.* Boca Raton, FL: Chapman & Hall/CRC Press. 2016. 361–386.

[97] Saddiki H, Balzer L. B. A Primer on Causality in Data Science. *Journal de la societe francaise de statistique.* 2020;161(1):67–90.

[98] van der Laan M. J, Rose S. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. New York, NY: Springer. 2018.

[99] Hernan M. A, Robins J. M. *Causal Inference: What If.* Boca Raton, FL: Chapman & Hall/CRC. 2020.

[100] Pearl J, Mackenzie D. *The book of why: the new science of cause and effect*. London, UK: Penguin Books. 2019.

[101] Lash T. L, Fox M. P, Fink A. K. *Applying quantitative bias analysis to epidemiologic data*. New York, NY: Springer. 2009.

[102] Rosenbaum P. R. *Observational studies*. New York, NY: Springer. 2nd ed. 2002.

[103] Rosenbaum P. R. *Design of observational studies*. New York, NY: Springer. 2010.

[104] Rubin D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology.* 1974;66(5):688–701.

[105] Rubin D. B. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine.* 2007;26(1):20–36.

[106] Cochran W. G, Chambers S. P. The Planning of Observational Studies of Human Populations. *Journal of the Royal Statistical Society. Series A (General).* 1965;128(2):234-266.

[107] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) . E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials. 2021. `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9r1-statistical-principles-clinical-trials-addendum-estimands-and-sensitivity-analysis-clinical`.

[108] Hernán M. A, Sauer B. C, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology.* 2016;79:70–75.

[109] Hernán M. A, Wang W, Leaf D. E. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA.* 2022;328(24):2446.

[110] Robins J. M, Hernán M. A. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. , eds. *Longitudinal Data Analysis.* Boca Raton, FL: Chapman & Hall/CRC Press. 2009. 553–597.

[111] Rubin D. B. [On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.] Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. *Statistical Science.* 1990;5(4):472-480.

[112] Pearl J. *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press. 2009.

[113] Holland P. W. Statistics and Causal Inference. *Journal of the American Statistical Association.* 1986;81(396):945–960.

[114] Pearl J. Causal diagrams for empirical research. *Biometrika.* 1995;82(4):669–688.

[115] Wright S. Correlation and Causation. *Journal of Agricultural Research.* 1921;20:557-585.

[116] Greenland S, Pearl J, Robins J. M. Causal diagrams for epidemiologic research. *Epidemiology.* 1999;10(1):37–48.

[117] Richardson T, Robins J. Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality. 2013. University of Washington Center for Statistics and the Social Sciences. `https://csss.uw.edu/Papers/wp128.pdf`.

[118] VanderWeele T. J. Principles of confounder selection. *European Journal of Epidemiology.* 2019;34(3):211–219.

[119] Stensrud M. J, Dukes O. Translating questions to estimands in randomized clinical trials with intercurrent events. *Statistics in Medicine.* 2022;41(16):3211–3228.

[120] Phillips A, Abellan-Andres J, Soren A, et al. Estimands: discussion points from the PSI estimands and sensitivity expert group. *Pharmaceutical Statistics.* 2017;16(1):6–11.

[121] Rudolph K. E, Gimbrone C, Matthay E. C, et al. When Effects Cannot be Estimated: Redefining Estimands to Understand the Effects of Naloxone Access Laws. *Epidemiology.* 2022;33(5):689–698.

[122] Gruber S, Phillips R. V, Lee H, Concato J, van der Laan M. Evaluating and improving real-world evidence with Targeted Learning. *arXiv.* Preprint posted online August 15, 2022. arXiv:2208.07283.

[123] Weber A. M, van der Laan M. J, Petersen M. L. Assumption Trade-Offs When Choosing Identification Strategies for Pre-Post Treatment Effect Estimation: An Illustration of a Community-Based Intervention in Madagascar. *Journal of Causal Inference.* 2015;3(1):109–130.

[124] Díaz I, van der Laan M. J. Sensitivity Analysis for Causal Inference under Unmeasured Confounding and Measurement Error Problems. *The International Journal of Biostatistics.* 2013;9(2):149-160.

[125] VanderWeele T. J, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine.* 2017;167(4):268-274.

[126] Robins J. M. Causal Inference from Complex Longitudinal Data. In: Bickel P, Diggle P, Fienberg S, et al. , eds. *Latent Variable Modeling and Applications to Causality.* New York, NY: Springer. 1997. 69–117.

[127] Shpitser I. Complete Identification Methods for the Causal Hierarchy. *Journal of Machine Learning Research.* 2008;9:1941–1979.

[128] Textor J, Zander B, Gilthorpe M. S, Liśkiewicz M, Ellison G. T. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology.* 2017;45(6):1887-1894.

[129] Sharma A, Kiciman E. DoWhy: An End-to-End Library for Causal Inference. *arXiv.* Preprint posted online November 9, 2020. arXiv:2011.04216.

[130] Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science.* 2001;16(3):199-231.

[131] Hernán M. A. The Hazards of Hazard Ratios. *Epidemiology.* 2010;21(1):13–15.

[132] Greenland S. Absence of Confounding Does Not Correspond to Collapsibility of the Rate Ratio or Rate Difference. *Epidemiology.* 1996;7(5):498–501.

[133] Martinussen T, Vansteelandt S, Andersen P. K. Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis.* 2020;26(4):833–855.

[134] Montoya L. M, Kosorok M. R, Geng E. H, Schwab J, Odeny T. A, Petersen M. L. Efficient and robust approaches for analysis of sequential multiple assignment randomized trials: Illustration using the ADAPT-R trial. *Biometrics.* 2022:1-15.

[135] U.S. Food and Drug Administration . Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry. 2019. `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry`.

[136] Gruber S, Lee H, Phillips R, Ho M, van der Laan M. Developing a Targeted Learning-Based Statistical Analysis Plan. *Statistics in Biopharmaceutical Research.* 2022;00(00):1–8.

[137] Greenland S. Basic Methods for Sensitivity Analysis of Biases. *International Journal of Epidemiology.* 1996;25(6):1107–1116.

[138] Lash T. L, Fox M. P, MacLehose R. F, Maldonado G, McCandless L. C, Greenland S. Good practices for quantitative bias analysis. *International Journal of Epidemiology.* 2014;43(6):1969–1985.

[139] Robins J. M, Rotnitzky A, Scharfstein D. O. Sensitivity Analysis for Selection bias and unmeasured Confounding in missing Data and Causal inference models. In: Miller W, Halloran M. E, Berry D. , eds. *Statistical Models in Epidemiology, the Environment, and Clinical Trials.* New York, NY: Springer. 2000. 1–94.

[140] Rotnitzky A, Robins J. M, Scharfstein D. O. Semiparametric Regression for Repeated Outcomes with Nonignorable Nonresponse. *Journal of the American Statistical Association.* 1998;93(444):1321–1339.

[141] Kasy M, Spiess J. Rationalizing Pre-Analysis Plans: Statistical Decisions Subject to Implementability. *arXiv.* Preprint posted online August 20, 2022. arXiv:2208.09638.

[142] Phillips C. V, LaPole L. M. Quantifying errors without random sampling. *BMC Medical Research Methodology.* 2003;3(1):9.

[143] Rosenbaum P. R, Rubin D. B. Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society: Series B (Methodological).* 1983;45(2):212–218.

[144] Cornfield J, Haenszel W, Hammond E. C, Lilienfeld A, Shimkin M, Wynder E. Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions. *JNCI: Journal of the National Cancer Institute.* 1959;22(1):173-203.

[145] Rudolph K. E, Keyes K. M. Voluntary Firearm Divestment and Suicide Risk: Real-World Importance in the Absence of Causal Identification. *Epidemiology.* 2023;34(1):107–110.

[146] Ford I, Norrie J. Pragmatic Trials. *New England Journal of Medicine.* 2016;375(5):454–463.

[147] Kim M, Harun N, Liu C, Khoury J. C, Broderick J. P. Bayesian selective response-adaptive design using the historical control. *Statistics in Medicine.* 2018;37(26):3709–3722.

[148] Ventz S, Khozin S, Louv B, et al. The design and evaluation of hybrid controlled trials that leverage external data and randomization. *Nature Communications.* 2022;13(1):5783.

[149] Chaudhuri S. E, Ho M. P, Irony T, Sheldon M, Lo A. W. Patient-centered clinical trials. *Drug Discovery Today.* 2018;23(2):395–401.

[150] U.S. Food and Drug Administration . Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products. 2020. `https://www.fda.gov/media/130897/download`.

[151] U.S. Food and Drug Administration . Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products. 2021. https://www.fda.gov/media/152503/download.

[152] U.S. Food and Drug Administration . Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products. 2021. https://www.fda.gov/media/154449/download.

[153] U.S. Food and Drug Administration . Data Standards for Drug and Biological Product Submissions Containing Real-World Data. 2021. https://www.fda.gov/media/153341/download.

[154] Robins J. A Graphical Approach to the Identification and Estimation of Causal Parameters in Mortality Studies with Sustained Exposure Periods. *Journal of Chronic Diseases.* ;40:139S–161S.

[155] Mahendraratnam N, Eckert J, Mercon K, et al. Understanding the Need for Non-Interventional Studies Using Secondary Data to Generate Real-World Evidence for Regulatory Decision Making, and Demonstrating their Credibility. 2019. Duke Margolis Center for Health Policy. https://healthpolicy.duke.edu/sites/default/files/2020-08/Non-Interventional%20Study%20Credibility.pdf.

[156] Benchimol E. I, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Medicine.* 2015;12(10):e1001885.

[157] Gatto N. M, Reynolds R. F, Campbell U. B. A Structured Preapproval and Postapproval Comparative Study Design Framework to Generate Valid and Transparent Real-World Evidence for Regulatory Decisions. *Clinical Pharmacology and Therapeutics.* 2019;106(1):103–115.

[158] Bykov K, Patorno E, D'Andrea E, et al. Prevalence of Avoidable and Bias-Inflicting Methodological Pitfalls in Real-World Studies of Medication Safety and Effectiveness. *Clinical Pharmacology and Therapeutics.* 2022;111(1):209–217.

[159] Rodbard H. W, Lingvay I, Reed J, et al. Semaglutide Added to Basal Insulin in Type 2 Diabetes (SUSTAIN 5): A Randomized, Controlled Trial. *The Journal of Clinical Endocrinology & Metabolism.* 2018;103(6):2291–2301.

[160] Ahrén B, Masmiquel L, Kumar H, et al. Efficacy and safety of once-weekly semaglutide versus once-daily sitagliptin as an add-on to metformin, thiazolidinediones, or both, in patients with type 2 diabetes (SUSTAIN 2): a 56-week, double-blind, phase 3a, randomised trial. *The Lancet Diabetes & Endocrinology.* 2017;5(5):341–354.

[161] Ahmann A. J, Capehorn M, Charpentier G, et al. Efficacy and Safety of Once-Weekly Semaglutide Versus Exenatide ER in Subjects With Type 2 Diabetes (SUSTAIN 3): A 56-Week, Open-Label, Randomized Clinical Trial. *Diabetes Care.* 2018;41(2):258–266.

[162] Aroda V. R, Bain S. C, Cariou B, et al. Efficacy and safety of once-weekly semaglutide versus once-daily insulin glargine as add-on to metformin (with or without sulfonylureas) in insulin-naive patients with type 2 diabetes (SUSTAIN 4): a randomised, open-label, parallel-group, multicentre, multinational, phase 3a trial. *The Lancet Diabetes & Endocrinology.* 2017;5(5):355–366.

[163] Marso S. P, Bain S. C, Consoli A, et al. Semaglutide and Cardiovascular Outcomes in Patients with Type 2 Diabetes. *New England Journal of Medicine.* 2016;375(19):1834–1844.

[164] Aroda V. R, Rosenstock J, Terauchi Y, et al. PIONEER 1: Randomized Clinical Trial of the Efficacy and Safety of Oral Semaglutide Monotherapy in Comparison With Placebo in Patients With Type 2 Diabetes. *Diabetes Care.* 2019;42(9):1724–1732.

[165] Rodbard H. W, Rosenstock J, Canani L. H, et al. Oral Semaglutide Versus Empagliflozin in Patients With Type 2 Diabetes Uncontrolled on Metformin: The PIONEER 2 Trial. *Diabetes Care.* 2019;42(12):2272–2281.

[166] Rosenstock J, Allison D, Birkenfeld A. L, et al. Effect of Additional Oral Semaglutide vs Sitagliptin on Glycated Hemoglobin in Adults With Type 2 Diabetes Uncontrolled With Metformin Alone or With Sulfonylurea: The PIONEER 3 Randomized Clinical Trial. *JAMA.* 2019;321(15):1466–1480.

[167] Pratley R, Amod A, Hoff S. T, et al. Oral semaglutide versus subcutaneous liraglutide and placebo in type 2 diabetes (PIONEER 4): a randomised, double-blind, phase 3a trial. *The Lancet.* 2019;394(10192):39–50.

[168] Husain M, Birkenfeld A. L, Donsmark M, et al. Oral Semaglutide and Cardiovascular Outcomes in Patients with Type 2 Diabetes. *New England Journal of Medicine.* 2019;381(9):841–851.

[169] McGuire D. K, Busui R. P, Deanfield J, et al. Effects of oral semaglutide on cardiovascular outcomes in individuals with type 2 diabetes and established atherosclerotic cardiovascular disease and/or chronic kidney disease: Design and baseline characteristics of SOUL, a randomized trial. *Diabetes Obesity and Metabolism.* 2023:dom.15058.

[170] Marso S. P, Poulter N. R, Nissen S. E, et al. Design of the liraglutide effect and action in diabetes: Evaluation of cardiovascular outcome results (LEADER) trial. *American Heart Journal.* 2013;166(5):823–830.e5.

[171] Bain S. C, Mosenzon O, Arechavaleta R, et al. Cardiovascular safety of oral semaglutide in patients with type 2 diabetes: Rationale, design and patient baseline characteristics for the PIONEER 6 trial. *Diabetes Obesity and Metabolism.* 2019;21(3):499–508.

[172] Buse J. B, Wexler D. J, Tsapas A, et al. 2019 Update to: Management of Hyperglycemia in Type 2 Diabetes, 2018. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care.* 2020;43(2):487–493.

[173] Hobbs B. P, Carlin B. P, Mandrekar S. J, Sargent D. J. Hierarchical Commensurate and Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials. *Biometrics.* 2011;67(3):1047–1056.

[174] Ghadessi M, Tang R, Zhou J, et al. A roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). *Orphanet Journal of Rare Diseases.* 2020;15(1):69.

[175] Chow C. J, Habermann E. B, Abraham A, et al. Does Enrollment in Cancer Trials Improve Survival?. *Journal of the American College of Surgeons.* 2013;216(4):774–780.

[176] U.S. Department of Health and Human Services . 45 - Public Welfare. 2018. https://www.govinfo.gov/app/details/CFR-2018-title45-vol1/CFR-2018-title45-vol1-sec164-514.

[177] Office for Civil Rights . Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Information Insurance Portability and Accountability Act (HIPAA) Privacy Rule. 2020. https://www.hhs.gov/guidance/document/guidance-regarding-methods-de-identification-protected-health-information-accordance-0.

[178] Franklin J. M, Patorno E, Desai R. J, et al. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. *Circulation.* 2021;143(10):1002-1013.

[179] Valentin G, Ravn M, Jensen E, et al. Socio-economic inequalities in fragility fracture incidence: a systematic review and meta-analysis of 61 observational studies. *Osteoporos International.* 2021;32(12):2433–2448.

[180] Yan X, Lee S, Li N. Missing Data Handling Methods in Medical Device Clinical Trials. *Journal of Biopharmaceutical Statistics.* 2009;19(6):1085–1098.

[181] Hobbs B. P, Carlin B. P, Sargent D. J. Adaptive adjustment of the randomization ratio using historical control data. *Clinical Trials.* 2013;10(3):430–440.

[182] The PREFER Consortium . PREFER Recommendations: Why, when and how to assess and use patient preferences in medical product decision-making. 2022. https://www.imi-prefer.eu/recommendations/.

[183] U.S. Food and Drug Administration . Type 2 Diabetes Mellitus: Evaluating the Safety of New Drugs for Improving Glycemic Control: Guidance for Industry. 2020. https://www.fda.gov/media/135936/download.

[184] R Core Team . R: A language and environment for statistical computing. 2022.

[185] Hastie T. *gam: Generalized Additive Models.* 2022. R package version 1.22.

Appendix A : Appendices for Chapter 1

A.1 Appendix A.1: Table of symbols

Table A.1: Table of Symbols

Symbol	**Meaning**
S	Variable indicating experiment (RCT or RCT+RWD)
A	Intervention
W	Covariates
Y	Outcome
X	Endogenous variables
U	Exogenous variables
O^n	Observed data $O^n = (O_1, ..., O_n)$
$P_{U,O}$	True distribution of full data (endogenous and exogenous variables)
P_0	True distribution of observed data
P_n	Empirical distribution of observed data
$P_{n,v}$	Empirical distribution of estimation set for cross-validation fold v
P_{n,v^c}	Empirical distribution of experiment-selection set (fold v)
$P_{0,n}$	True data distribution dependent on n
$P^*_{n,v}$	Indicates training of initial estimators for outcome regression and treatment mechanism on experiment-selection set for fold v with TMLE targeting on separate or pooled estimation set(s)
P^*_{n,v^c}	Indicates training and TMLE targeting of estimator for outcome regression on experiment-selection set for fold v

$Q^{\{0,s\}}$	$E[Y\|S \in \{0,s\}, A, W]$
Q^{s}	$E[Y\|S \in \{0,s\}, S, A, W]$
Q^{NCO}	$E[NCO\|S \in \{0,s\}, A, W]$
$Q^*_{n,v}$	Indicates updated Q after training of initial estimators for outcome regression Q_{n,v^c} and treatment mechanism on experiment-selection set for fold v with TMLE targeting on separate or pooled estimation set(s)
Q^*_{n,v^c}	Indicates updated Q after training and TMLE targeting of outcome regression on experiment-selection set for fold v
g^s	$P(S = 0\|S \in \{0,s\}, A = 0, W)$ (**"Selection Mechanism"**)
g^a	$P(A = a\|S \in \{0,s\}, W)$ (**"Treatment Mechanism"**)
$\Psi^F_s(P_{U,O})$	$E_{W\|S\in\{0,s\}}[E(Y^1 - Y^0\|W, S \in \{0,s\})]$ (**"ATE"**)
$\Psi_s(P_0)$	$E_{W\|S\in\{0,s\}}[E_0[Y\|A = 1, S \in \{0,s\}, W] - E_0[Y\|A = 0, S \in \{0,s\}, W]]$
$\tilde{\Psi}^F_s(P_{U,O})$	$E_{W\|S\in\{0,s\}}[E(Y^{a=1,s=0} - Y^{a=0,s=0}\|W, S \in \{0,s\})]$ (**"ATE-RCT"**)
$\tilde{\Psi}_s(P_0)$	$E_{W\|S\in\{0,s\}}[E_0[Y\|A = 1, S = 0, W] - E_0[Y\|A = 0, S = 0, W]]$
$\Phi_s(P_0)$	$E_{W\|S\in\{0,s\}}[E_0[NCO\|A = 1, S \in \{0,s\}, W] - E_0[NCO\|A = 0, S \in \{0,s\}, W]]$
$s^\star_n$	$\underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + (\hat{\Psi}^\#_s(P_n))^2$ (Bias^2 + variance selector **"b2v"**)
$s^{\star\star}_n$	$\underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + (\hat{\Psi}^\#_s(P_n) + \hat{\Phi}_s(P_n))^2$ (Selector including ATE on NCO **"+nco"**)
$s^{\star\star\star}_n$	$\underset{s}{argmin} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + (\hat{\Phi}_s(P_n))^2$ (Bias only estimated as ATE on NCO **"nco only"**)

NCO: Negative control outcome. ATE: Average treatment effect. ATE-RCT: Average treatment effect if participants were in RCT.

A.2 Appendix A.2: Estimation of bias

In order to estimate the bias, $\Psi_s^{\#}(P_0)$, we will use targeted maximum likelihood estimation [40]. The efficient influence curve (EIC) for

$$\Psi_s^0(P_0) = E_{W|S\in\{0,s\}}[E_0[Y|A=0,S\in\{0,s\},W]] \text{ is}$$
$$D^*_{\Psi_s^0}(O) = \frac{I(S\in\{0,s\})}{P_0(S\in\{0,s\})}(\frac{I(A=0)}{g_0^a(A=0|S\in\{0,s\},W)}(Y - Q_0^{\{0,s\}}(S\in\{0,s\},A,W)) + Q_{s,0}^{\{0,s\}}(S\in \{0,s\},0,W) - \Psi_s^0(P_0))$$

where $g_0^a(A=0|S\in\{0,s\},W)$ is the probability that $A=0$ given $S\in\{0,s\}$, W, and $Q_0^{\{0,s\}}(S\in\{0,s\},A,W) = E_0[Y|S\in\{0,s\},A,W]$.

TMLE involves fitting initial estimates of the treatment mechanism, g_n^a, and outcome regression, $Q_n^{\{0,s\}}$, with the SuperLearner ensemble machine learning algorithm [61]. The initial estimate is then targeted using a parametric working model [41]:

$$\text{logit}(Q_n^{\{0,s\},*}(S\in\{0,s\},A,W)) = \text{logit}(Q_n^{\{0,s\}}(S\in\{0,s\},A,W)) + \epsilon_n H^*_{s,n}(S\in\{0,s\},A,W)$$

where $H^*_{s,n}(S\in\{0,s\},A,W) = \frac{I(A=0,S\in\{0,s\})}{g_n^a(A=0|S\in\{0,s\},W)P_n(S\in\{0,s\})}$ is the covariate in front of the residual in the EIC, and ϵ_n may be fitted using logistic regression of Y on $H^*_{s,n}$ with offset $\text{logit}(Q_n^{\{0,s\}}(S\in\{0,s\},A,W))$ among observations with $S\in\{0,s\}$. While a linear regression may be performed for a continuous outcome, it is common practice to scale the outcome as $(Y-min(Y))/(max(Y)-min(Y)))$, perform the TMLE with a logistic fluctuation, and re-scale the parameter estimate to the original scale in order to respect the bounds of the observed data distribution [69].

The conditional expectation of the counterfactual outcome under control is updated as:

$$\text{logit}(Q_n^{\{0,s\},*}(S\in\{0,s\},0,W)) = \text{logit}(Q_n^{\{0,s\}}(S\in\{0,s\},0,W)) + \epsilon_n H^*_{s,n}(S\in\{0,s\},0,W)$$

where $H^*_{s,n}(S\in\{0,s\},0,W) = \frac{1}{g_n^a(A=0|S\in\{0,s\},W)P_n(S\in\{0,s\})}$ using the same ϵ_n. The final estimate of the mean outcome under control in the combined dataset is then

$$\hat{\Psi}_s^0(P_n) = \frac{1}{n}\sum_{i=1}^n \frac{I(S_i\in\{0,s\})}{P(S\in\{0,s\})} Q_n^{\{0,s\},*}(S_i\in\{0,s\},0,W_i)$$

An alternate option that may be more stable in the context of near-violations of the positivity assumption is to move the denominator of the clever covariate to the denominator of the weights for the regression training the TMLE coefficient [42, 43, 44]. For this option of "targeting the weights", we perform a logistic regression of binary or scaled-continuous Y on $H^*_{s,n}(S\in\{0,s\},A,W) = I(A=0,S\in\{0,s\})$ with offset $logit(Q_n^{\{0,s\}}(S\in\{0,s\},A,W))$ and weights $\frac{I(A=0,S\in\{0,s\})}{g_n^a(A=0|S\in\{0,s\},W)P_n(S\in\{0,s\})}$ among observations with $S\in\{0,s\}$. Initial estimates are then updated as

$$Q_n^{\{0,s\},*}(S\in\{0,s\},0,W) = logit^{-1}(logit(Q_n^{\{0,s\}}(S\in\{0,s\},0,W)) + \epsilon_n).$$

We can also use TMLE to estimate $\tilde{\Psi}_s^0(P_0) = E_{W|S\in\{0,s\}}[E_0[Y|A=0,S=0,W]]$, with EIC

$$D^*_{\tilde{\Psi}_s^0}(O) = \frac{I(S\in\{0,s\})}{P_0(S\in\{0,s\})}(\frac{I(S=0,A=0)}{g_0^s(S=0|A=0,S\in\{0,s\},W)g_0^a(A=0|S\in\{0,s\},W)}(Y - Q_0^s(S\in \{0,s\},A,S,W)) + Q_0^s(S\in\{0,s\},0,0,W) - \tilde{\Psi}_s^0(P_0))$$

where $Q_0^s(S \in \{0,s\}, A, S, W) = E_0[Y|S \in \{0,s\}, A, S, W]$. We use the same procedure as above with clever covariate
$H^*_{s,n}(S \in \{0,s\}, A, S, W) = \frac{I(S=0,A=0)}{g_n^s(S=0|A=0,S\in\{0,s\},W=w)g_n^a(A=0|S\in\{0,s\},W=w)P_n(S\in\{0,s\})}$ to obtain a targeted estimate $Q_n^{s,*}(S \in \{0,s\}, 0, 0, W)$. Our updated estimate
$\hat{\bar{\Psi}}_s^0(P_n) = \frac{1}{n}\sum_{i=1}^n \frac{I(S_i\in\{0,s\})}{P_n(S\in\{0,s\})}Q_n^{s,*}(S_i \in \{0,s\}, 0, 0, W_i)$. Then, our TMLE estimate of the bias

$$\hat{\Psi}_s^{\#}(P_n) = \hat{\bar{\Psi}}_s^0(P_n) - \hat{\Psi}_s^0(P_n)$$

A.3 Appendix A.3: Proof of Theorem 1

Proof of Theorem 1:

$$\sqrt{n}(\psi_n - \psi_{n,0}) = \frac{\sqrt{n}}{V}\sum_{v=1}^{V}(\hat{\Psi}_{s_n^\star(v^c)}(P^*_{n,v}) - \Psi_{s_n^\star(v^c)}(P_0)) =$$
$$\frac{\sqrt{n}}{V}\sum_{v=1}^{V}((P_{n,v} - P_0)D^*_{s_n^\star(v^c)}(P^*_{n,v}) + R_{s_n^\star(v^c)}(P^*_{n,v}, P_0))$$
$$= \frac{\sqrt{n}}{V}\sum_{v=1}^{V}((P_{n,v}-P_0)D^*_{s_n^\star(v^c)}(P_0)+(P_{n,v}-P_0)(D^*_{s_n^\star(v^c)}(P^*_{n,v})-D^*_{s_n^\star(v^c)}(P_0))+R_{s_n^\star(v^c)}(P^*_{n,v}, P_0))$$
$$= \frac{\sqrt{n}}{V}\sum_{v=1}^{V}((P_{n,v} - P_0)D^*_{s_n^\star(v^c)}(P_0)) + o_P(1)$$

by assumption of *Conditions 1 and 2*. Define

$$Z_n^\dagger(s, v) = \sqrt{n}(P_{n,v} - P_0)D^*_s(P_0)$$

By the Central Limit Theorem, across all s and v, the vector $Z_n^\dagger = (Z_n^\dagger(s, v) : s, v) \sim N(\vec{0}, \Sigma^\Psi)$.

In order to understand the behavior of $\sqrt{n}(\psi_n - \psi_{n,0})$, we must also understand the behavior of $s_n^\star(v^c)$, which depends on the behavior of either $Z_n^\#(s, v^c)$ or $Z_n^{\#+\Phi}(s, v^c)$.

For the standardized bias terms estimated on experiment-selection sets,

$$Z_n^\#(s, v^c) = \sqrt{n}(\hat{\Psi}_s^\#(P^*_{n,v^c}) - \Psi_s^\#(P_0))$$
$$= \sqrt{n}((P_{n,v^c} - P_0)D^*_{\Psi_s^\#,v^c}(P^*_{n,v^c}) + R_{\#,s}(P^*_{n,v^c}, P_0))$$
$$= \sqrt{n}((P_{n,v^c} - P_0)D^*_{\Psi_s^\#,v^c}(P_0) + (P_{n,v^c} - P_0)\{D^*_{\Psi_s^\#,v^c}(P^*_{n,v^c}) - D^*_{\Psi_s^\#,v^c}(P_0)\} + R_{\#,s}(P^*_{n,v^c}, P_0))$$
$$= \sqrt{n}(P_{n,v^c} - P_0)D^*_{\Psi_s^\#,v^c}(P_0) + o_P(1)$$

by assumption of *Conditions 1, 2, and 3*. If $\hat{\Phi}_s$ is include in the bias estimation, where $D^*_{(\#+\Phi)_s,v^c} = D^*_{\Psi_s^\#,v^c} + D^*_{\Phi_s,v^c}$,

$$Z_n^{\#+\Phi}(s, v^c) = \sqrt{n}((\hat{\Psi}_s^\# + \hat{\Phi}_s)(P^*_{n,v^c}) - (\Psi_s^\# + \Phi_s)(P_0))$$
$$= \sqrt{n}((P_{n,v^c} - P_0)D^*_{(\#+\Phi)_s,v^c}(P^*_{n,v^c}) + R_{(\#+\Phi),s}(P^*_{n,v^c}, P_0))$$
$$= \sqrt{n}((P_{n,v^c} - P_0)D^*_{(\#+\Phi)_s,v^c}(P_0) + (P_{n,v^c} - P_0)\{D^*_{(\#+\Phi)_s,v}(P^*_{n,v^c}) - D^*_{(\#+\Phi)_s,v^c}(P_0)\} +$$
$$R_{(\#+\Phi),s}(P^*_{n,v^c}, P_0))$$
$$= \sqrt{n}(P_{n,v^c} - P_0)D^*_{(\#+\Phi)_s,v^c}(P_0) + o_P(1)$$

by assumption of *Conditions 1, 2, and 3*.

By the Central Limit Theorem, $Z_n^\#(s, v^c)$ and $Z_n^{\#+\Phi}(s, v^c)$ also converge to normal distributions. Across all s and v,

$$Z_n^\# = (Z_n^\#(s, v^c) : s, v) \sim N(\vec{0}, \Sigma^\#)$$
$$Z_n^{\#+\Phi} = (Z_n^{\#+\Phi}(s, v^c) : s, v) \sim N(\vec{0}, \Sigma^{\#+\Phi})$$
$$\tilde{Z} = (Z^\#, Z^\dagger) \sim N(\vec{0}, \tilde{\Sigma})$$
$$\text{or } \tilde{Z} = (Z^{\#+\Phi}, Z^\dagger) \sim N(\vec{0}, \tilde{\Sigma})$$

where $\tilde{\Sigma}$ is defined in Chapter 1, Section 1.5.1. The limit distribution of the experiment-selector CV-TMLE is then defined by sampling from $\tilde{Z}$, calculating

$$\bar{s}^{\star}(v^c) = \underset{s}{argmin}\ \sigma^2_{D^*_{\Psi_{s,v^c}}} + (Z^{\#}(s,v^c) + \sqrt{n}\Psi^{\#}_s(P_0))^2$$

or

$$\bar{s}^{\star\star}(v^c) = \underset{s}{argmin}\ \sigma^2_{D^*_{\Psi_{s,v^c}}} + (Z^{\#+\Phi}(s,v^c) + \sqrt{n}(\Psi^{\#}_s(P_0) + \Phi_s(P_0)))^2$$

and finally calculating

$$\sqrt{n}(\psi_n - \psi_{n,0}) = \frac{\sqrt{n}}{V}\sum_{v=1}^{V}(P_{n,v} - P_0)D^*_{\Psi_{\bar{s}^{\star}(v^c)}}(P_0) + o_P(1) = \frac{1}{V}\sum_{v=1}^{V}(Z^{\dagger}(\bar{s}^{\star}(v^c), v)) + o_P(1)$$

or

$$\sqrt{n}(\psi_n - \psi_{n,0}) = \frac{\sqrt{n}}{V}\sum_{v=1}^{V}(P_{n,v} - P_0)D^*_{\Psi_{\bar{s}^{\star\star}(v^c)}}(P_0) + o_P(1) = \frac{1}{V}\sum_{v=1}^{V}(Z^{\dagger}(\bar{s}^{\star\star}(v^c), v)) + o_P(1)$$

$\sqrt{n}(\psi_n - \psi_{n,0})$ thus converges to a mixture of normal distributions.

A.4 Appendix A.4: Data generating process for simulation

As described in Chapter 1, Section 1.6.1, four datasets were simulated as follows: 1) an "RCT" dataset of 150 observations with $S = 0$, $A = 1$ randomized with probability 0.67, and bias terms B_1 and B_2 equal to zero, 2) a "real-world" dataset of 500 observations with $S = 1$, $A = 0$, and bias terms B_1 and B_2 equal to zero, 3) "RWD" of 500 observations with $S = 2$, $A = 0$, and $B_1 + B_2 \approx B$, and 4) "RWD" of 500 observations with $S = 3$, $A = 0$, and $B_1 + B_2 \approx 5 * B$. For this simulation, biased RWD could be included if it is approximately $\sqrt{\frac{\hat{\sigma}^2_{D^*_{\Psi_{s=0,n,v^c}}}}{n} - \frac{\hat{\sigma}^2_{D^*_{\Psi_{s\in\{0,2\},n,v^c}}}}{n}} = B = 0.21$. We generate B_1 and B_2 as described below:

Dataset	B_1	B_2
$S = 0$	0	0
$S = 1$	0	0
$S = 2$	$N(\frac{3}{4}B, 0.02^2)$	$N(\frac{1}{4}B, 0.02^2)$
$S = 3$	$N(\frac{3}{4} * 5 * B, 0.02^2)$	$N(\frac{1}{4} * 5 * B, 0.02^2)$

We also simulate two covariates, W1 and W2, as $N(0, 1)$. The outcome Y and NCO are then simulated as

$$Y = -3 + 2 * W1 + W2 - 0.6 * A + B_1 + B_2 + U_Y$$
$$NCO = -2 + W1 + 2 * W2 + B_1 + U_{nco}$$

with $U_Y \sim N(0, 1.5^2)$ and $U_{nco} \sim N(0, 1.5^2)$. The true causal ATE of A on Y in this simulation is -0.6. The NCO is affected by B_1 but not B_2, and so the U-comparability and additive equi-confounding assumptions do not completely hold in this case.

For the TMLE-based methods used in the simulation, we use linear regression for the outcome regression and a candidate library for the treatment mechanism consisting of lasso regression [54] or the mean. We use 10-fold cross-validation and target the weights as described above. When only $S = 0$ data is considered, we use the true randomization probability for $g(A|W)$.

Appendix B : Appendices for Chapter 4

B.1 Appendix B.1: Mathematical notation for causal and statistical estimands

As described in Table 4.1, the causal question of interest is: what would the difference in risk of MACE (defined as death from any cause, nonfatal MI, or nonfatal stroke) within one year be if all patients in a population consistent with the PIONEER 6 inclusion/exclusion criteria and timeframe[168], and with similar healthcare engagement, were prescribed oral semaglutide plus standard-of-care compared to if all patients were prescribed standard-of-care alone, and if censoring had been prevented for all patients?

The two intervention variables that are modified in our treatment strategies are A – an indicator of prescribing patients oral semaglutide in addition to standard-of-care (A=1) or standard-of-care (A=0) – and C – an indicator of whether the participants were censored before one year. We denote our outcome of MACE by 1-year as Y. Because some participants are censored, the observed outcome, $Y^{\star}$, is the true value of MACE for those who were not censored and whose outcomes were measured correctly and missing for those who were censored.

We then define the following potential outcomes[28, 111]; $Y^{a=1,c=0}$ is the one-year MACE status an individual would have had if they had been prescribed oral semaglutide in addition to standard-of-care and not been censored, and $Y^{a=0,c=0}$ is the one-year MACE status an individual would have had if they had been prescribed standard-of-care and not been censored. The simplest mathematical representation of a causal estimand that answers our question is given by the causal risk difference:

$$E(Y^{a=1,c=0} - Y^{a=0,c=0}).$$

Note that this causal risk difference is defined with respect to a specific target population. Despite efforts to ensure comparability between the RCT population and external control RWD, our approach acknowledges that the RCT and RWD populations may nonetheless have different distributions of baseline characteristics. Because the proposed estimator (ES-CVTMLE) only augments the control arm when the RWD meet pre-specified criteria (evaluated across multiple internal sample splits), the exact target population to which the causal risk difference applies will depend on the extent to which these criteria are met and the RCT standard-of-care arm is augmented with RWD.

More formally, let S be a variable describing study participation, where S=0 indicates that an individual participated in an RCT and S=1 indicates that an individual participated in the real-world healthcare system. Designs 1 and 2 only utilize RCT data, and so in these designs, we can only evaluate the causal risk difference within the RCT context and for a target population represented by the RCT participants. We can rewrite the causal parameter to represent the causal risk difference

(not adjusted for baseline characteristics) in a way that makes explicit that it refers to the RCT context:

$$\psi^{\star}_{RCT,unadj} = E(Y^{a=1,c=0} - Y^{a=0,c=0}|S = 0).$$

With baseline covariates, W, the adjusted causal risk difference for the RCT context and target population is:

$$\psi^{\star}_{RCT,adj} = E_{W|S=0}[E(Y^{a=1,c=0} - Y^{a=0,c=0}|W, S = 0)].$$

We note that the true causal risk difference in the non-inferiority and superiority trials could be different if they had different inclusion and exclusion criteria or if there were changes over time in the background standard-of-care. For simplicity, however, we will consider that the non-inferiority and superiority RCTs target the same causal parameter, $\psi^{\star}_{RCT,unadj}$.

In hybrid Design 3, we consider integrating extra RWD controls with our non-inferiority trial and only run the superiority RCT if the null hypothesis of the superiority RCT is not rejected in the hybrid analysis. The ES-CVTMLE adjusts for baseline covariates whether the RWD is included or rejected, so if the hybrid design rejects the RWD, analyzing the non-inferiority RCT only, then the causal target parameter is $\psi^{\star}_{RCT,adj}$.

In contrast, if the hybrid design does select to augment the non-inferiority trial with extra RWD controls, this may modify the target population if the RWD controls have a different (though overlapping) distribution of baseline covariates compared to the RCT population. Inclusion of RWD controls may also modify the target parameter if the true effect of oral semaglutide versus standard-of-care is different in the RCT and RWD contexts. The causal risk difference in the combined RCT plus RWD experiment that integrates S=0 and S=1 is given by:

$$\psi^{\star}_{RCT,RWD} = E_{W|S\in\{0,1\}}[E(Y^{a=1,c=0} - Y^{a=0,c=0}|W, S \in \{0, 1\})].$$

In the hybrid design, we use the data from the real-world source and the non-inferiority trial to decide whether to estimate the causal risk difference for the RCT context and target population ($\psi^{\star}_{RCT,adj}$) or the causal risk difference for the hybrid RCT-RWD context and target population ($\psi^{\star}_{RCT,RWD}$), where either parameter represents an answer to our question for that particular population and context.

We will ultimately use the experiment-selector CV-TMLE [60] to analyze the results of the hybrid trial. This method uses cross-validation to separate the part of the data that is used to choose whether to attempt to estimate $\psi^{\star}_{RCT,adj}$ or $\psi^{\star}_{RCT,RWD}$ and the part of the data that is used for estimation of the corresponding risk difference. The decision of whether to augment the RCT with external control data may differ in different cross-validation folds. The causal estimand would then be interpreted as the causal risk difference for a target population that is a weighted average of the RCT population and external control RWD population. More formally, let the target parameter chosen for a given fold, v, be $\psi^{\star}_{v}$. The overall causal target parameter for the hybrid design is then the average of the causal target parameters selected in each fold. For example, with ten cross-validation folds, the causal target parameter would be

$$\psi^{\star}_{hybrid} = \frac{1}{10}\sum_{v=1}^{10}\psi^{\star}_{v}.$$

Please see Dang et al. (2022) [60] for further details of this methodology.

Using the g-formula [29], we may define the statistical estimands (functions of the observed data) that are as close as possible to the causal effects of interest for each study design, where

$$\psi_{RCT,unadj} = E(Y^\star|C=0,A=1,S=0) - E(Y^\star|C=0,A=0,S=0)$$
$$\psi_{RCT,adj} = E_{W|S=0}[E(Y^\star|C=0,A=1,W,S=0) - E(Y^\star|C=0,A=0,W,S=0)]$$
$$\psi_{RCT,RWD} = E_{W|S\in\{0,1\}}[E(Y^\star|C=0,A=1,W,S\in\{0,1\}) -$$
$$E(Y^\star|C=0,A=0,W,S\in\{0,1\})]$$

ψ_v is whichever of $\psi_{RCT,adj}$ or $\psi_{RCT,RWD}$ was selected in cross-validation fold v, and

$$\psi_{hybrid} = \tfrac{1}{10}\sum_{v=1}^{10}\psi_v.$$

B.2 Appendix B.2: Assessment of plausibility of causal identification assumptions

First, we consider whether the causal effect is identified in Designs 1 and 2 based on the directed acyclic graph (DAG) in Figure 4.2a, using the backdoor criterion[112]. Because treatment was randomized, there are no unmeasured common causes of treatment and the outcome. Furthermore, because censoring is minimal in the RCTs (it was truly 0.3% by one year in PIONEER 6)[168], the magnitude of bias that could result from potential unmeasured common causes of censoring and MACE is likely to be negligible. For these reasons, we expect to identify the causal effect of interest using Designs 1 and 2.

As discussed in the main text, Design 3 does not assume that the causal effect of interest is identified in the pooled RCT and RWD. However, taking steps to improve the plausibility of causal identification assumptions for the combined data also increases the likelihood that RWD will be integrated in the hybrid design. Again using the backdoor criterion[112] and our DAG in Figure 4.2b, we consider possible reasons for a causal gap in an analysis of the pooled RCT and RWD. First, there would be a causal gap if being in the RCT versus the real world affected outcomes outside of the effect due to prescribing either oral semaglutide or standard-of-care; this is certainly possible for the reasons described in Step 2 above.

If we were able to conduct a pragmatic clinical trial for the randomized component of our hybrid design in which the trial aimed to mimic real-world care as closely as possible outside of baseline treatment randomization, then it would be more likely that trial participation only affected outcomes through treatment assignment[174, 34]. In this case, however, we are not able to consider a pragmatic RCT because the current FDA draft guidance on "Evaluating the Safety of New Drugs for Improving Glycemic Control"[183] requires a sufficient number of phase 3 clinical trial patient-years on the medication of interest during which time CV outcomes are evaluated by adjudication to evaluate cardiovascular safety. Instead, we may attempt to select RWD controls who at least have similar healthcare access and engagement compared to RCT participants (discussed below).

There would also be a causal gap for the pooled RCT and RWD analysis if there were unmeasured common causes of trial participation and MACE or of censoring and MACE, and we expect a larger amount of censoring in the real world compared to the RCT. We apply the same inclusion and exclusion criteria and timeframe for the RWD and RCT controls, yet the question remains whether

the measured baseline characteristics that are indicative of demographics, baseline health status, and treatment are sufficient to adjust for common causes of our intervention variables and outcome.

To try to minimize the amount of bias that would be introduced by integrating RCT and RWD controls, we consider a further restriction of the CDM cohort to select patients who are likely to be at a similar disease stage with similar healthcare access and engagement compared to the RCT participants. Selecting RWD patients prescribed DPP4is is one method of making disease stage and engagement comparable[178]. We also exclude CDM patients with missingness in the baseline covariates, expecting that patients for whom this laboratory and medical history data is not recorded might not be followed as closely by their providers as patients in the RCT.

Additionally, there would be a causal gap if assignment to standard-of-care in the trial and the RWD were not equivalent in terms of their effects on MACE. The most obvious reason this might not be true is that participants in the RCT were prescribed an inactive placebo pill and were not prescribed a DPP4i based on exclusion criteria, while the RWD participants were prescribed a DPP4i as an active comparator. The question then is whether the effect of being assigned placebo is different from the effect of being prescribed a DPP4i on the outcome of MACE. In the RCT Duplicate study, DPP4is were chosen as a "proxy for placebo" relative to the outcome of MACE in studies of GLP-1RAs "because they are antidiabetic treatments that have similar indications to the treatments under study, but they are not known to have any effect on the cardiovascular outcomes of interest based on recent evidence"[178]. If this reasoning is correct, assignment to placebo should have the same effect as assignment to a DPP4i for the primary outcome. Another question is whether the background standard-of-care that patients receive is equivalent in the RCT and the RWD. While it is possible that there are differences between the standard-of-care provided by trial versus non-trial clinicians, we attempted to ensure that "standard-of-care" would be as similar as possible by restricting the CDM cohort to the same time period as PIONEER 6.

Finally, to identify a causal effect in the combined RCT and RWD, we need sufficient data support. In other words, participants in any stratum of measured confounders must have a positive probability of being assigned to either intervention strategy: oral semaglutide and not being censored or standard-of-care and not being censored. This assumption is also known as the positivity assumption[30, 31]. Because we only add extra RWD controls, including any RWD participant whose particular combination of measured potential confounding variables was not shared by RCT participants would violate the positivity assumption. We solve this problem by limiting the CDM cohort to participants whose baseline covariates were within the range of baseline covariates represented in the trial population.

B.3 Appendix B.3: Estimation of the causal gap

The first estimate of the causal gap used by the ES-CVTMLE compares conditional mean outcomes between RCT and combined RCT-RWD controls. The statistical estimand for this causal gap parameter is given by

$$\Psi^{\#} = E_{W|S\in\{0,1\}}[E[Y^{\star}|C=0,A=0,S=0,W] - E[Y^{\star}|C=0,A=0,S\in\{0,1\},W]].$$

The ES-CVTMLE estimates $\Psi^{\#}$ using targeted maximum likelihood estimation [41, 40], but the precision of the estimate depends on the sample size of the RCT. In a given sample dataset, the estimate of $\Psi^{\#}$ will not be exactly equal to the true causal gap because of finite sample variabil-

ity. Nonetheless, $\Psi^{\#}$ represents our best estimate of the causal bias that would be introduced by including RWD controls in the analysis. See Dang et al. (2022) [60] for more details.

We also estimate the causal gap as the estimated average treatment effect on a negative control outcome (NCO). NCOs are not affected by the treatment but ideally should be affected by as many of the factors that lead to violations of identification assumptions as possible [24]. Any estimated association between treatment and the NCO is thus due either to a causal gap or due to finite sample variability.

B.4 Appendix B.4: Simulation set-up

The data for the simulation were generated as follows. First, we generate data to mimic a non-inferiority RCT (RCT1) of sample size n1=3183, twenty-one different "real-world" datasets (RWD) of sample size n2=2483, and a superiority RCT (RCT2) of sample size n3=9500. In the two "RCT" datasets, treatment is randomized with probability 0.5. In the "RWD", all participants receive A=0. Two baseline covariates, W1 and W2 are drawn from $Normal(\mu = 0, \sigma = 1)$ distributions for participants from all studies.

We generate 21 potential levels of bias in the "RWD" as follows. B is a variable that introduces bias when non-zero. The value of B is zero for the two "RCT" datasets and the unbiased "RWD" dataset. For the remaining 20 "RWD" datasets, the value of B ranges from positive 1/10*0.65 to 10/10*0.65 in increments of 1/10*0.65 and from (-1)/10*1.7 to (-10)/10*1.7 in increments of (-1)/10*1.7. These values were chosen because, due to the properties of the $logit^{-1}$ function, this range of values of B leads to true bias as large as $\pm 2.1\%$. This maximum magnitude of bias in either direction was chosen so that the true bias minus two times the standard error of the bias estimator would be larger than the standard error of the risk difference TMLE estimator for the RCT alone. Because the ES-CVTMLE will select the combination of RCT and RWD if the estimated squared bias plus the variance of the TMLE risk difference estimator for the combined data is smaller than the estimated squared bias plus the variance of the TMLE risk difference estimator for the RCT alone, these magnitudes of bias include the full spectrum of magnitudes for which we would expect that RWD might be included in the analysis in some simulation iterations.

The primary outcome, Y, is generated as follows:

$$Y \sim Bernoulli(p = logit^{-1}(-3.33 + 0.2 * W1 - 0.4 * W2 + U_y + B))$$

where $U_y \sim Normal(\mu = 0, \sigma = 0.5)$. This equation was designed so that the overall probability of MACE would be similar to the true probability of MACE in the PIONEER 6 placebo arm (4.2%). Adding extra random error, U_y, means that the baseline covariates are not very predictive of the outcome, which is common with relatively rare binary outcomes measured years after baseline.

In order that the magnitude of the effect of B on the relationship between the treatment and the negative control outcome (NCO) be similar to the magnitude of the effect of B on the relationship between the treatment and the true outcome, but to make sure that the primary and negative control outcomes are not too tightly correlated, we let

$$NCO \sim Bernoulli(p = logit^{-1}(-3.33 + 0.2 * W1 - 0.4 * W2 + U_{nco} + B))$$

where $U_{nco} \sim Normal(\mu = 0, \sigma = 0.5)$ but is independent of U_y. The simulation in Appendix B.5 describes an alternate, "worst-case" simulation in which B has no effect on the NCO.

We also generate some missing outcomes, where the indicator that the outcome is censored ($C = 1$) and the indicator that the NCO is censored ($C_{nco} = 1$) are generated as follows:

$$C \sim Bernoulli(p = (1 - logit^{-1}(2.2 + W1 - W2 + 4.5 * I(S = 0))))$$
$$C_{nco} \sim Bernoulli(p = (1 - logit^{-1}(2.2 + W1 - W2 + 4.5 * I(S = 0))))$$

where S=0 indicates one of the simulated RCT datasets. These equations for outcome missingness were designed to approximate the true rates of outcome missingness in the RCT context (0.3% for PIONEER 6) and the RWD context (16% for CDM).

Note that in this simulation, treatment, A, does not affect the outcome, Y. If treatment were to affect the outcome, then the true causal risk difference (CRD) would be different for different values of B, even if there were no interaction term between B and A, due to the properties of the $logit^{-1}$ function. Instead, because A does not affect Y, the true value of the CRD for all combinations of RCT and RWD is zero. This sets up an even competition between the study designs when different RWD are included in the hybrid analysis; we would expect the power to reject the same null hypothesis to depend on bias and variance but not on different true causal effects when different combinations of data are analyzed. We report 95% CI coverage for the true causal risk difference of zero for all designs across the 1000 iterations of this simulation.

We also aim to evaluate the amount of person-time that participants are precluded from receiving a GLP1-RA because they are in the control arm of one of the potential RCTs. As shown in Supplementary Table B.1, for Designs 1 and 3, we start by determining whether the results of RCT1 or of the hybrid RCT1-RWD analysis reject the null hypothesis. Because the simulated effect is zero, and we expect a truly negative effect of semaglutide versus standard-of-care on MACE based on the results of the SUSTAIN 6 trial[163], we shift the definition of the null hypothesis to be that the risk difference is a specified value larger than zero.

For the sake of this demonstration, we consider a significant result as an estimate of the risk difference with an upper 95% CI limit less than positive 1.1%. This value was chosen because a trial of 9500 participants (similar to our simulated RCT2) would be expected to have power of 0.8 to detect a risk difference of -1.1% (using $\alpha = 0.05$). If we view the simulated data as having simulated CRD values that are shifted 1.1% more positive than the value of the CRD that our superiority trial would be powered to detect, then shifting standard criteria for superiority by the same amount would cause us to conclude that a result was significant if the estimated upper bound on the 95% confidence interval were less than positive 1.1%. Power is then calculated as the proportion of iterations in which this modified null hypothesis is rejected (i.e., the proportion of iterations in which the 95% CI did not include 1.1%).

Note that this method of evaluating significance should actually be slightly conservative given that in the simulated data, the probability of the outcome is approximately the PIONEER6 placebo arm rate (4.2%) in both arms, whereas if a negative risk difference had been simulated, the treatment arm outcome probability would have been less than 4.2%. For example, with the same sample size N, if the treatment arm probability of the outcome were 3%, the variance of the difference in sample proportions (V1) would be smaller than the variance if the treatment arm probability of the outcome were also 4.2% (V2):

$$V1 \approx \frac{0.03(1 - 0.03)}{N} + \frac{0.042(1 - 0.042)}{N} = \frac{0.069}{N}$$

$$< \frac{0.080}{N} = \frac{0.042(1-0.042)}{N} + \frac{0.042(1-0.042)}{N} \approx V2$$

Finally, the person-time participants are prevented from receiving any GLP1-RA for each design in a single iteration of the simulation is calculated as described in Supplementary Table B.1 below. We report the average amount of person-time during which participants were prevented from receiving a GLP1-RA across all iterations for each design. While the event-driven SOUL trial will actually run for closer to four years, we only include person-time required to evaluate the outcome for this proposed study: MACE by one year after baseline.

Supplementary Table B.1: Calculation of Person-Time prevented from Receiving a GLP1-RA for each Design

1	1. If RCT1 result is significant: 1 year x 1591.5 placebo arm participants = 1591.5 person-years 2. If RCT1 result not significant: 1 year x 1591.5 placebo arm participants from RCT1 + 1 year x 4750 placebo arm participants from RCT2 = 6341.5 person-years
2	1. 1 year x 4750 placebo arm participants from RCT2 = 4750 person-years
3	1. If hybrid RCT1-RWD result is significant†: 1 year x 1591.5 placebo arm participants from RCT 1 = 1591.5 person-years 2. If hybrid RCT1-RWD result not significant: 1 year x 1591.5 placebo arm participants from RCT1 + 1 year x 4750 placebo arm participants from RCT2 = 6341.5 person-years

†RWD participants were not prevented from receiving a GLP1-RA by being in an RCT control arm and so are not included in the amount of person-time during which patients are prevented from receiving a GLP1-RA.

A simple Super Learner[61] library of candidate algorithms was used for the ES-CVTMLE to improve computational efficiency of this simulation. The outcome regression was estimated using logistic regression. The Super Learner for the censoring mechanism ($P(C = 0|A, W)$) and treatment mechanism ($P(A = 1|W)$) for the combined RCT and RWD considered either a logistic regression or the sample mean. Because missingness was negligible in the RCT, the missingness mechanism in the RCT used within the ES-CVTMLE estimator only considered the sample proportion of non-missing outcomes. R Statistical Software version 4.2.2 was used for all simulations [184].

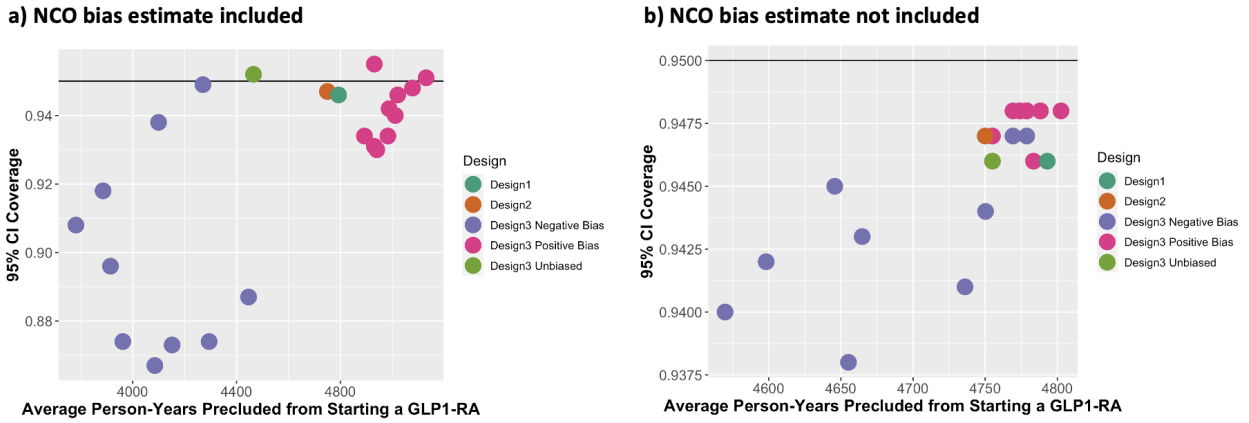
B.5 Appendix B.5: Data generation and results for simulation in which bias has no effect on NCO

We also include a simulation in which the bias term, B, has no effect on the NCO. This simulation is included as a worst-case scenario for how hybrid Design 3 could perform under a complete violation of the assumption that the factors causing bias in the relationship between the treatment and the true outcome also cause bias in the relationship between the treatment and the NCO. The process for generating the data for this simulation is the same as in Appendix B.4, except that

$$NCO \sim Bernoulli(p = logit^{-1}(-3.33 + 0.2 * W1 - 0.4 * W2 + U_{nco}))$$

Supplementary Figure B.1 shows the results of 1000 iterations of this simulation both when the ES-CVTMLE uses the estimated average treatment effect on the NCO as an estimate of the causal gap and when the ES-CVTMLE only estimates bias based on the method described in Appendix B.3, without considering the NCO.

Supplementary Figure B.1: Simulation Results by Study Design with Different Amounts of RWD Bias when Bias has No Effect on NCO



Because the bias term, B, has no effect on the NCO in this simulation, hybrid Design 3 is more likely to incorporate biased RWD in either direction when the NCO is used to estimate bias. With bias in the positive direction (towards the null), the hybrid RCT-RWD design is less likely to reject the null hypothesis, leading to the follow-up RCT being run in more iterations. As a result, patient-time during which participants were precluded from receiving a GLP1-RA was larger on average for Design 3 with positive bias than for Designs 1 and 2.

With bias in the negative direction, Design 3 led to as large as an average of 1012 fewer person-years during which participants could not start a GLP1-RA compared to Design 1 but had 95% CI coverage ranging from 0.867 to 0.949. These results demonstrate the importance of choosing a good negative control outcome if this particular study design and estimator are selected.

If an appropriate negative control outcome is not available, the ES-CVTMLE may also only use one estimate of bias, as described in Appendix B.3. In this context, the ES-CVTMLE is more

conservative (less likely to include RWD), with coverage not less than 0.938 for any magnitude of bias, but a maximum average decrease in patient-years during which a GLP1-RA may not be prescribed of 223. The possibilities described in Supplementary Figure B.1 should also be considered in the process of study design and estimator selection. Also note that the performance of Design 1 compared to Design 2 is slightly different in this simulation compared to the simulation in the main text. Such variability is expected when the same simulation is run different times although may be decreased by running more iterations.

B.6 Appendix B.6: Further details regarding specification of the ES-CVTMLE and unadjusted estimators

The ES-CVTMLE estimator for the real data analysis used twenty cross-validation folds. The Super Learner libraries for the relevant regressions consisted of logistic regression or the sample mean for the outcome, logistic regression [53], a general additive model [185], or multivariate adaptive regression splines [55] for the propensity score, logistic regression [53], a general additive model [185], multivariate adaptive regression splines [55] or the sample mean for the RWD outcome missingness indicator, and the sample mean only for the RCT outcome missingness indicator. Because estimates vary somewhat when different random seeds are used to define cross-validation folds and train machine learning algorithms, the ES-CVTMLE estimator was run ten times with different random seeds, and the point estimate and upper and lower confidence interval bounds were averaged across these ten iterations.

Because censoring was negligible (0.3%) in the simulated and real data RCTs, a complete case analysis was conducted for the components of each study design that only involved RCT data.