

UC Berkeley

UC Berkeley Previously Published Works

Title

The Causal Roadmap and Simulations to Improve the Rigor and Reproducibility of Real-data Applications.

Permalink

<https://escholarship.org/uc/item/5nc7w75m>

Journal

Epidemiology, 35(6)

Authors

Nance, Nerissa

Petersen, Maya

van der Laan, Mark

et al.

Publication Date

2024-08-01

DOI

10.1097/EDE.0000000000001773

Peer reviewed

The Causal Roadmap and Simulations to Improve the Rigor and Reproducibility of Real-data Applications

 Nerissa Nance,^a Maya L. Petersen,^a Mark van der Laan,^a and Laura B. Balzer^a

Abstract: The Causal Roadmap outlines a systematic approach to asking and answering questions of cause and effect: define the quantity of interest, evaluate needed assumptions, conduct statistical estimation, and carefully interpret results. To protect research integrity, it is essential that the algorithm for statistical estimation and inference be prespecified prior to conducting any effectiveness analyses. However, it is often unclear which algorithm will perform optimally for the real-data application. Instead, there is a temptation to simply implement one's favorite algorithm, recycling prior code or relying on the default settings of a computing package. Here, we call for the use of simulations that realistically reflect the application, including key characteristics such as strong confounding and dependent or missing outcomes, to objectively compare candidate estimators and facilitate full specification of the statistical analysis plan. Such simulations are informed by the Causal Roadmap and conducted after data collection but prior to effect estimation. We illustrate with two worked examples. First, in an observational longitudinal study, we use outcome-blind simulations to inform nuisance parameter estimation and variance estimation for longitudinal targeted minimum loss-based estimation. Second, in a cluster randomized trial with missing outcomes, we use treatment-blind simulations to examine type-I error control in two-stage targeted minimum loss-based estimation. In both examples, realistic simulations empower us to prespecify an estimation approach with strong expected finite sample performance, and also produce quality-controlled computing code for the actual analysis. Together, this process helps to improve the rigor and reproducibility of our research.

Keywords: Causal inference; Causal Roadmap; Prespecification; Real-world data; Simulations; TMLE

(*Epidemiology* 2024;35: 791–800)


Formal frameworks for causal and statistical inference can help researchers to clearly structure and understand the links between their research question, causal model, data, statistical estimation, and results interpretation. Examples of such frameworks include the Causal Roadmap and target trial emulation.^{1–3} Recent commentaries on epidemiologic training have highlighted the role of such frameworks in asking thoughtful and feasible study questions, particularly amid a proliferation of novel analytic methods that may aid or distract from answering that question.^{4,5} Even after we have specified a well-defined and relevant question, there are many steps to setting up the analysis to answer it. For example, the remaining steps of the Causal Roadmap (hereafter, “the Roadmap”) are to: (2) specify a causal model reflecting background knowledge and uncertainties; (3) define the causal effect of interest; (4) describe the data available to answer the question; (5) assess identifiability; (6) select a statistical model and estimand; (7) estimate and obtain inference, and (8) interpret results.

Roadmap steps one to six set up a statistical estimation problem, reflecting our research question and the real-world challenges of the data. Specifically, the Roadmap leads us to a well-defined statistical estimand, which is a function of the observed data distribution, and a realistic statistical model. (Formally, the statistical model is the set of all possible observed data distributions.²) However, the Roadmap does not tell us which algorithm to apply for estimation and inference. While an algorithm's theoretical properties can narrow the scope of possibilities, it is often unclear a priori which approach will perform best in the real-data application. Instead, there is a tendency to simply apply one's preferred algorithm, using the default settings of a computing package or recycling prior code. Likewise, there is a temptation to try several implementations and pick the implementation that yields the most favorable or logical result. As detailed below, we advocate for the use of realistic simulations to objectively select the algorithm

Submitted February 27, 2024; accepted July 19, 2024

From the ^aUniversity of California, Berkeley, Berkeley, CA. Supported, in part, by a philanthropic gift from the Novo Nordisk corporation to the University of California, Berkeley for the Joint Initiative for Causal Inference (JICI). The funders had no role in the conceptualization or writing of the manuscript.

Disclosure: The authors report no conflicts of interest.

 Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Nerissa Nance and Laura B. Balzer, University of California, Berkeley, 2121 Berkeley Way, Berkeley, CA 94720. E-mail: nerissanance@berkeley.edu; lbalzer@berkeley.edu.

Copyright © 2024 The Author(s). Published by Wolters Kluwer Health, Inc.

This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

ISSN: 1044-3983/24/356-791800

DOI: 10.1097/EDE.0000000000001773

for estimation and inference and prespecify the statistical analysis plan.

The statistical analysis plan delineates key features of the real-data analysis, including the target population, primary outcome, exposure conditions, causal effect of interest, approaches to handling potential inferential threats (e.g., confounding and missing data), statistical estimand, primary/secondary analyses, and sensitivity analyses. While many of these features follow from earlier steps of the Roadmap, the statistical analysis plan requires us to state the precise implementation of the estimator, including approaches for estimating nuisance parameters and approaches for obtaining inference. (Nuisance parameters are quantities needed to evaluate the statistical estimand, but are not the estimand itself.) Full prespecification requires more than simply stating the statistical estimand (e.g., the longitudinal G-computation formula) and the general class of estimators (e.g., targeted minimum loss-based estimation [TMLE]). Indeed, the process of prespecifying a statistical analysis plan requires us to think critically about different estimation and inferential strategies as well as their expected performance before running any analyses to assess causality. As a result, prespecification of the statistical analysis plan helps improve transparency and protect against ad hoc analyses, which can lead to a “fishing expedition” to find the most promising results and inflated type-I error rates.

Regulatory and funding agencies typically require the statistical analysis plan to be prespecified prior to conducting effectiveness analyses in randomized trials.^{6–9} There is also a growing movement to improve the reproducibility and transparency of observational studies through rigorous planning and reporting.^{10–13} Mathur and Fox¹⁴ provide an excellent review of the principles and practices to improve open and reproducible research in epidemiology; in particular, they highlight preregistration of statistical analysis plans for observational studies and code sharing. Notably, Gruber et al.¹⁵ discuss how the Roadmap for Targeted Learning can inform the development of statistical analysis plans. Here, we build on this work by providing guidance and context on how to select the approach for statistical estimation and inference for the real-data application.

Our goal is to describe how finite sample simulations, informed by the Roadmap and reflecting the real-data application, can be used to objectively compare estimation strategies and develop a completely prespecified statistical analysis plan. For demonstration, we provide two worked examples: (1) an observational study with a time-varying exposure and censoring and (2) a randomized trial with missing and dependent outcomes. We also highlight how this approach naturally leads to fully prespecified and quality-checked computing code. Thus, our approach has the potential to improve the transparency, reproducibility, and rigor of our analyses aiming to evaluate causal effects.

Our presentation assumes familiarity with foundational concepts in causal and statistical inference (e.g., causal

models, identifiability assumptions, and the G-computation formula). For a review of these concepts and an introduction to the Causal Roadmap, we refer to Petersen and van der Laan¹ and Dang et al.¹⁶ An overview of the Roadmap for the running examples is provided in Table 1. Indeed, our worked examples are inspired by real studies and are inherently complex, highlighting the real-world challenges that commonly arise when aiming to infer causality. The remainder of the article is organized as follows. First, we outline how simulations are used in epidemiology. Then, we demonstrate the utility of the Roadmap in setting up the statistical estimation problem and designing the simulation study. Next, we describe how to conduct simulations for estimator selection in real-data applications; specifically, we discuss the prespecification of the candidate estimators, data-generating process, performance metrics, and selection scheme. Finally, we describe the consequences of our approach to improve research transparency and reproducibility.

ON SIMULATIONS IN EPIDEMIOLOGY

Simulation studies are widely applied in methodologic research to evaluate the finite sample properties of existing and recently developed estimators.²³ Since the true value of the target parameter is known, simulations enable us to calculate performance metrics, such as bias and confidence interval coverage. For example, the well-known Kang and Schafer censored data simulations revealed the instability of estimating equation-based methods under data sparsity and inspired suspicion of doubly robust approaches.²⁴ Subsequent replication of these simulations has highlighted the potential for doubly robust, substitution estimators (e.g., TMLE and collaborative TMLE) to overcome these challenges.²⁵ More recently, simulations have been used to illustrate the potential advantages and perils of using machine learning in analyses seeking to infer causality.^{26–28}

Beyond methods evaluation, epidemiology uses simulations in several other settings. Examples include teaching epidemiologic concepts, evaluating study designs, forecasting disease trajectories, agent-based modeling, addressing transportability, and data pooling.^{29–35} Prior to data collection, simulation studies are commonly implemented to inform the design randomized trials, including power calculations.^{36–38} Following data collection and after-effect estimation, simulations are also applied in sensitivity analyses, including quantitative bias analysis.^{39,40} However, to the best of our knowledge, there are few published examples of using simulations in the principled comparison and selection of estimators for a real-data analysis after data have been collected but before any effectiveness analyses are conducted. Some exceptions include the use of outcome-blind simulations to select the primary analysis in a SMART trial, to evaluate propensity score estimators within TMLE in a drug safety monitoring study, and to compare estimators for longitudinal effects with registry data.^{41–43}

TABLE 1. Overview of the Causal Roadmap for the Two Running Examples

| Causal Roadmap Steps | Observational Study Example | Randomized Trial Example |
|---|---|--|
| (1) Research question Specify the primary exposure(s), outcome, and target population | What is the effect of SGLT2 ^a inhibitor use on the risk of renal disease onset among patients with diabetes in an integrated healthcare system in the United States? | What is the effect of the multicomponent, SEARCH-Youth intervention on viral suppression among 15–24 year olds with HIV in rural Kenya and Uganda? ¹⁷ |
| (2) Causal model Describe the confounding structure, missing data mechanism, (in)dependence structure, ... | Longitudinal model including time-varying confounders and factors influencing censoring. | Multilevel model reflecting that the intervention is randomized at the clinic level, but outcomes are at the individual level and subject to missingness. |
| (3) Causal effect Using counterfactuals, specify the target effect and scale of interest | Causal risk difference: $\mathbb{E}[Y_{\bar{a}=1, \bar{c}=0}(t)] - \mathbb{E}[Y_{\bar{a}=0, \bar{c}=0}(t)]$ with $Y_{\bar{a}, \bar{c}}(t)$ as the counterfactual outcome at time t under SLGT2 use $\bar{A} = \bar{a}$ and no censoring $\bar{C} = 0$ throughout follow-up. | Sample prevalence ratio: $\frac{1/N \sum_{i=1}^N Y_i^c(1)}{1/N \sum_{i=1}^N Y_i^c(0)}$ with $Y_i^c(\alpha)$ as the counterfactual % with viral suppression in clinic $i = \{1, \dots, N\}$ under treatment level $\mathcal{A}^c = \alpha^c$ and complete outcome measurement. |
| (4) Observed data Describe the data that have been or will be observed, include the exposure(s), outcome(s), ... | Longitudinal data structure: history of covariates $\bar{L}(t)$, exposure $\bar{A}(t)$, censoring $\bar{C}(t)$, and outcomes $\bar{Y}(t)$ through time t . | Multilevel data structure: cluster- and individual-level baseline covariates ($\mathcal{L}^c, \mathcal{W}$), cluster-level exposure \mathcal{A}^c , and individual-level, postbaseline covariates \mathcal{M} , measurement indicator Δ , and outcome Y . |
| (5) Identifiability Evaluate the assumptions needed to express the causal effect as statistical estimand | No unmeasured confounding, sufficient data support (i.e., no practical positivity violations), and censoring at random | Within each clinic and values of adjustment variables, individual-level outcomes are missing at random and there is a positive probability of measurement |
| (6) Define the statistical estimation problem Specify the statistical estimand and model | The statistical estimand is the iterated conditional expectation expression of the longitudinal g-formula. ^{18,19} The statistical model is nonparametric. ^b | Within each clinic, the cluster-level endpoint, accounting for missingness, is $Y^c = \mathbb{E}[\mathbb{E}[Y \Delta = 1, \mathcal{W}, \mathcal{M}]]$. The statistical estimand to evaluate the intervention effect is $\Psi = \frac{1/N \sum_{i=1}^N \mathbb{E}[Y_i^c \mathcal{A}^c = 1, \mathcal{L}_i^c]}{1/N \sum_{i=1}^N \mathbb{E}[Y_i^c \mathcal{A}^c = 0, \mathcal{L}_i^c]}$. ²⁰ The statistical model is semiparametric. ^b |
| (7) Estimate and obtain inference In a prespecified way, chose and implement an estimator | Outcome-blind simulations to objectively select between alternative implementations of longitudinal TMLE ^c ; see Table 2. ^{21,22} | Treatment-blind simulations to objectively select between alternative implementations of two-stage TMLE ^c ; see Table 2. ²⁰ |
| (8) Interpret State the results in light of the causal and statistical assumptions | After flexibly accounting for measured time-varying confounders, 2 years of continuous use of SGLT2 ^a inhibitors was associated with a 5% (95% CI: 2.75%, 7.25%) decrease in the risk of renal disease onset. | After flexibly accounting for missing outcomes and clustering, SEARCH-Youth increased viral suppression among adolescents and young adults with HIV by 10% (risk ratio=1.10, 95% CI: 1.03, 1.16). ¹⁷ |

^aSodium-glucose cotransporter 2 inhibitors.
^bSee the Statistical Analysis Plan in the eAppendices; <http://imks.lww.com/EDE/C169> for additional details.
^cTargeted minimum loss-based estimation.

A pertinent commentary in the *British Medical Journal* called for the broader use of simulations to inform applied data analyses, but also recognized that the implementation and reporting of such studies is the subject of continued debate.⁴⁴ We aim to help address these and other issues by guiding researchers on the use simulations, informed by the Roadmap and reflecting the real-data application, to aid in the development and full prespecification of the statistical analysis plan and corresponding computing code.

DEFINING THE ESTIMATION PROBLEM WITH THE CAUSAL ROADMAP

As illustrated in Table 1, the first six steps of the Roadmap setup the statistical estimation problem, which is defined by the statistical estimand and the statistical model.¹ Of course, one could specify these elements without the Roadmap. In our experience, however, applying the Roadmap has several strengths relative to other frameworks and the following benefits.⁵ Among others, the Roadmap helps clarify the research goals, highlight potential inferential threats, specify the handling of events occurring after the initial exposure or treatment, and facilitate transparent discussions about the plausibility of assumptions. Perhaps most crucially, the Roadmap leads to a statistical estimand reflecting our original research question as well as the real-world challenges in the data. In other words, even if the identifiability assumptions do not hold, the Roadmap guides us to statistical estimand coming as close as possible to the wished-for effect. (The size of the “causal gap” can be formally explored in sensitivity analyses and is taken into account during interpretation.⁵) In most cases, our statistical estimand is a complicated function of the observed data distribution and not equal to a single coefficient in a parametric regression. This complexity is needed to generate the most appropriate answer to our research question and often precludes the use of more traditional statistical approaches. Equally important, the Roadmap highlights that we rarely have the knowledge to support functional form assumptions, beyond treatment randomization in a trial. Instead, our statistical model is often nonparametric or semiparametric, and we need to harness machine learning during estimation to avoid unsubstantiated assumptions.

As a concrete example, consider a study aiming to evaluate the effect of a time-varying and nonrandomized exposure: sustained use of sodium-glucose cotransporter 2 (SGLT2) inhibitors on the onset of renal disease among patients with diabetes. As shown in Table 1, application of the Roadmap highlights the potential for bias and misleading inference due to confounding, censoring, and practical violations of the positivity assumption, occurring when there is insufficient variability in the exposure within confounder strata.^{45,46} These inferential threats can be particularly fraught in settings with longitudinal exposures; the longer follow-up time, the more potential there is for time-dependent confounding, right-censoring, and lower support for the longitudinal

exposures of interest. Given these challenges, the Roadmap leads to a complex statistical estimand: a contrast of the iterated conditional expectation expression of the longitudinal G-computation formula (eAppendix A; <http://links.lww.com/EDE/C169>).^{18,19} Importantly, the Roadmap also leads to a nonparametric statistical model without functional form assumptions. Altogether, the Roadmap narrows the scope of possible estimators to algorithms that can handle time-dependent confounding, right-censoring, and poor data support as well as harness machine learning to avoid unsubstantiated modeling assumptions. For this setting, common approaches include singly robust estimators, such as inverse probability weighting and G-computation, as well as doubly robust alternatives, such as augmented inverse probability weighting and TMLE.^{2,18,47–49} Each has statistical properties that may lend themselves (or not) to a specific analysis. As described below, we can use simulations, informed by the Roadmap, to choose the estimator expected to perform best in the actual analysis.

As a second example, consider the SEARCH-Youth study, a cluster randomized trial to evaluate the effect of a multicomponent intervention on viral suppression among youth with HIV in East Africa.¹⁷ As shown in Table 1, the Roadmap highlights the impacts of randomizing the treatment to health clinics (instead of individuals) and missing data. Specifically, each Roadmap step reflects the dependence between participants within clinics and the potential biases from the missing data, equivalent to time-dependent confounding. Again, the Roadmap leads to a complex statistical estimand: a contrast of clinical-level summary measures, each accounting for baseline and postbaseline causes of measurement and outcomes (eAppendix B; <http://links.lww.com/EDE/C169>).^{20,50,51} The Roadmap also leads us to a semiparametric statistical model, only reflecting our knowledge of treatment randomization. Here, the Roadmap narrows the set of possible estimators to those that flexibly handle dependent and missing data, specifically, approaches allowing the missingness mechanism to vary by cluster.²⁰ In two-stage TMLE, for example, we first estimate a summary measure accounting for missing data in cluster separately and then evaluate the intervention effect on those cluster-level summaries. Additionally, as common in cluster randomized trials,⁵² few clinics were randomized, specifically 28, in SEARCH-Youth. Therefore, we also need an estimation and inferential approach that performs well with few independent units. Again, simulations can aid in the formal evaluation of alternatives and prespecification of the primary analysis.

SIMULATIONS TO INFORM THE REAL-DATA ANALYSIS

We now detail how simulations, informed by the Roadmap and reflecting the real-data application, can aid in objectively selecting and appropriately implementing the estimation and inferential approach expected to perform best in the real-data analysis. To do so, we need to prespecify the

candidate estimators, data-generating process for the simulations, performance metrics, and selection process.

Choosing the Candidate Estimators

It is essential to choose candidates targeting the statistical estimand of interest. This may seem obvious, but without careful consideration, we could end up comparing estimators of marginal versus conditional effects, especially in hierarchical data settings.^{22,50,53,54} As previously discussed, following the Roadmap narrows the set of candidate algorithms to those targeting the statistical estimand. This set can further be narrowed by considering the asymptotic properties of the estimators (e.g., efficiency, double robustness). Even if we settle on a single class of estimators, such as TMLE, there are still many decisions before the statistical analysis plan is fully specified.

We must decide how to estimate nuisance parameters. In doubly robust estimators, for example, nuisance parameters typically include the outcome regressions (i.e., the conditional expectation of the outcome given past exposure/measurement and covariates) and propensity scores (i.e., the conditional probability of exposure/measurement given the past). To respect our statistical model, machine learning is often required for flexible, data-adaptive estimation of nuisance parameters. However, the application of machine learning requires additional choices. For example, in the ensemble algorithm Super Learner, we need to specify the candidate learners (including their tuning parameters), the cross-validation scheme, and the loss function.^{55,56} After obtaining initial estimates of the nuisance parameters, there may be additional decisions. For example, with practical positivity violations, we can decide to truncate the estimated propensity scores at various levels.^{45,57} Finally, there are a variety of options for statistical inference. For TMLE, for example, some approaches for variance estimation are the nonparametric bootstrap, standard or cross-validated estimates of influence curve, plug-in estimation of the variance, or other doubly robust options.^{2,21,58–60}

For our running examples, Table 2 provides an overview of the candidate approaches that were prespecified for objective comparison in simulations. In the observational study, the candidate algorithms were longitudinal TMLE with various implementations. For nuisance parameter estimation, Super Learner with and without covariate screening and bounded or unbounded estimates of the propensity score were considered. For statistical inference, candidates included Wald-Type 95% confidence intervals with variance estimated by the influence curve or the nonparametric bootstrap. For the cluster randomized trial, we limited the candidate algorithms to two-stage TMLE with the following specifications. For estimation of the cluster-level endpoints accounting for missing outcomes, candidates were TMLE using Super Learner, TMLE using parametric regressions, and the empirical mean among those measured. For estimation of the intervention effect, candidates were TMLE with various approaches to covariate adjustment for precision gains.^{58,61} Finally, candidates for variance

estimation included standard or cross-validated estimates of the influence curve. We now discuss how to define the data-generating process for the simulation to formally evaluate the performance of these candidates.

Defining the Data Generation Process for the Simulation

Thus far, the Roadmap has aided in defining the statistical estimation problem and specifying the set of candidate algorithms for estimation and inference. Simulations reflecting the real-data application can facilitate objective comparison and selection between these candidates if we choose a data-generating process that is close to the real one. Concretely, the application of the Roadmap highlighted several potential biases and inferential threats in the running examples (Table 1). For the observational study, we need to design a simulation with, at minimum, the same exposure/confounder/censoring structure and, therefore, the same practical positivity challenges as the real data. For the randomized trial, we need to design a simulation with, at minimum, the same number of clusters, a similar distribution of participants per cluster, and a plausible missing data mechanism as the real data. Given these specifications, several options exist.

Monte Carlo simulations, where we repeatedly sample from a known data-generating process, are common and traditionally employ parametric models for data generation.⁶³ Such parametric models often fail to reflect the complexities of the real data, especially in longitudinal or clustered data settings. Considering the limitations of fully parametric simulations, plasmode simulations have gained popularity and may be particularly useful for our focus: estimator selection after data have been collected but before effect estimation. Plasmode simulations, as defined here, encompass a range of semiparametric methods that sample partially from the empirical data distribution, while allowing for some user specification.^{45,64}

There are various types of plasmode simulations. In “outcome-blind” plasmode simulations, we preserve the relationships between the baseline covariates, while simulating the outcome (and other variables) through parametric or semiparametric methods.^{41,42,65,66} In these simulations, the value of the (simulated) effect is known, but we remain blinded to the true exposure–outcome relationship. As described in Table 2 and eAppendix A; <http://links.lww.com/EDE/C169>, outcome-blind simulations were conducted in the observational study by resampling the baseline covariates from the empirical distribution and then applying highly adaptive least absolute shrinkage and selection operator to simulate the longitudinal exposures, censoring, time-varying covariates, and outcome.⁶⁷ This approach preserves the complex relationships between baseline covariates while generating the remaining variables to reflect challenges in the real-data application (e.g., poor data support due to the rare exposure, long-term follow-up, and strong confounding).

TABLE 2. Overview of Simulation Setup and Results for the Two Running Examples

| Simulation Specifications and Results | Observational Study Example | Randomized Trial Example |
|---|--|---|
| (1) Statistical estimation problem ^a | The longitudinal G-computation formula under practical positivity violations with a long-term rare exposure, right-censoring, and rare outcome + non-parametric statistical model. | Two-stage estimand with (i) differential missingness of individual-level outcomes; (ii) few clusters randomized, and (iii) the cluster as the independent unit + semiparametric statistical model. |
| (2) Estimators compared ^b | Longitudinal TMLE ^c with alternative approaches for (i) nuisance parameter estimation using Super Learner with/without covariate prescreening and with/without truncation of the estimated propensity scores; (ii) variance estimation: influence curve vs. nonparametric bootstrap. | Two-stage TMLE ^c with alternative approaches to (i) account for missing individual-level outcomes: the empirical mean, TMLE with main terms regression, or TMLE with Super Learner; (ii) adaptively adjust to improve precision: adaptive prespecification with a limited vs. expanded adjustment set; |
| (3) Data generation process | Outcome-blind simulations preserving the baseline covariate structure, while simulating exposure-censoring variables with positivity challenges and a synthetic outcome with a similar marginal distribution as the real data. | (iii) obtain inference with standard or cross-validated estimates of the influence curve. ^{20,58,61} Treatment-blind simulations preserving the covariate-outcome data structure but randomly permuting the treatment indicator + generation of outcome measurement indicators by an independent statistician. |
| (4) Performance metrics and selection approach ^b | Over 1000 iterations, (i) select the approach for nuisance parameter estimation that minimizes empirical variance and preserves Oracle coverage; (ii) given (i), select the approach for variance estimation that minimizes the estimated variance and preserves 95% confidence interval coverage. | Over 1000 iterations, (i) select the approach for the cluster-level endpoints resulting in nominal confidence interval coverage for those endpoints and type-I error control for the overall effect; (ii) given (i), select the approach for effect estimation and variance estimation resulting in optimal type-I error control. |
| (5) Resulting primary analysis ^b | Longitudinal TMLE ^c with Super Learner with algorithm prescreening and without propensity score truncation, with influence curve-based variance estimation. ^{21,22,62} | Two-stage TMLE using (i) TMLE ^c with Super Learner to flexibly estimate the cluster-level endpoints, (ii) TMLE with adaptive prespecification with limited candidates to efficiently estimate the intervention effect; (iii) variance estimation with the cluster-level, influence curve. |

^aSee Table 1 for details.

^bSee the SAPs in the eAppendices; <http://links.lww.com/EDE/C169> for details, including the precise descriptions of the candidate estimators (e.g., the library, loss function, and cross-validation scheme for Super Learner), performance metrics, and the primary analysis.

^cTargeted minimum loss-based estimation.

“Treatment-blind” simulations are another plasmode simulation technique where the covariate-outcome data are preserved but the treatment indicator is randomly permuted.⁶¹ As detailed below, treatment-blind simulations are particularly relevant for evaluating type-I error control, because the null hypothesis is true by design. As outlined in Table 2 and eAppendix B; <http://links.lww.com/EDE/C169>, such simulations were implemented in the trial example by randomly shuffling the treatment indicator and imposing missingness on outcomes through a measurement indicator, which was generated by an independent statistician and as a function of the baseline cluster-level and individual-level covariates, the permuted treatment indicator, and time-varying covariates. This simulation approach preserves the covariates and underlying outcomes, while facilitating a rigorous comparison of alternative approaches and their potential to reduce bias due to differential outcome measurement and improve efficiency through covariate adjustment, as described next.

Specifying the Performance Metrics and Selection Approach

Once we have the set of candidate estimators and data-generating process for the simulations, we need to prespecify the performance metrics and process to objectively compare the candidates. In Table 3, we review some common metrics, such as the bias and variance of the point estimates as well as 95% confidence interval coverage (i.e., the proportion of calculated confidence intervals that contain the true effect). To compare estimators in a way that is agnostic to the variance estimator and evaluate the extent to which an estimator’s bias is negligible, we can use “Oracle coverage,” where the 95% confidence intervals are calculated using the variance of the point estimates across the simulation iterations, instead of the estimated variance. In simulations to inform randomized trials, common metrics include statistical power (i.e., the proportion of times the false null hypothesis is rejected) and type-I error control (i.e., the proportion of times the true null hypothesis is rejected). We may additionally be interested in estimating the potential savings in sample size to achieve the same power.^{61,65,68} Finally, we prespecify the selection process for objectively choosing the best-performing candidate and, thereby, the primary analytic approach.

For the running examples, Table 2 provides the performance metrics, selection process, and final estimator. In both studies, selection was a two-step process, implemented in R, and with 1000 simulation iterations. In the observational study, the optimal approach for nuisance parameter estimation was first selected to minimize the empirical variance but preserve Oracle coverage. Then given this choice, the optimal approach for inference was selected to minimize the variance estimate but preserve 95% confidence interval coverage. In the randomized trial, the optimal approach for estimating the cluster-level endpoints accounting for missing outcomes was based on attaining nominal confidence interval coverage for

TABLE 3. Examples of Performance Metrics to Use in Simulations to Inform Real-data Analyses

| Performance Metric | Calculation | Relevance |
|--------------------------------------|--|--|
| Bias (point estimates) | The average difference between the point estimate and the target effect across the simulation iterations | What is the accuracy of the estimator? |
| Variance (point estimates) | The variance of the point estimates across the simulation iterations | How precise is the estimator? |
| Mean-squared error | The average of the squared differences between the point estimate and target effect (equivalent to Bias ² + variance) | What is the variability of the estimator around the target effect? ⁶⁹ (Akin to asking how is the estimator balancing bias and variance?) |
| Bias-variance ratio | Ratio between the bias and variance | Is the estimator’s bias disappearing at a fast enough rate relative to its variance? |
| Variance to estimated variance ratio | Ratio between the variance of the point estimates and the average variance estimate | Is the variance being over- or underestimated? |
| Oracle coverage | The proportion of 95% confidence intervals, calculated with the variance of the point estimates, containing the target effect | Is the bias in the point estimates impacting the estimator’s confidence interval coverage? |
| Confidence interval coverage | The proportion of 95% confidence intervals, each calculated with the estimated variance, containing the target effect | Does the approach for estimation and inference result in valid inference? |
| Power | The proportion of simulation iterations where the true null hypothesis is rejected | Will the approach for estimation and inference identify an effect when it exists? |
| Type-I error | The proportion of simulation iterations where the true null hypothesis is rejected | Will the approach for estimation and inference lead to the incorrect conclusion of an effect when none exists? |

those endpoints and type-I error control for the intervention effect. Then given this choice, the optimal approach for estimation and inference for the intervention effect was selected to maximize precision without sacrificing type-I error. Estimation approaches with good, but not optimal, performance were then prespecified as sensitivity analyses.

FOSTERING TRANSPARENT AND REPRODUCIBLE RESEARCH

Informed by the Roadmap, we have conducted a simulation study, reflecting the real-data application and facilitating objective comparison of various approaches for estimation and inference. Specifically, we prespecified our candidates, the data-generating process, the performance measures, and the selection scheme. With this simulation study, we have a responsible and “hands-off” approach to selecting the best estimator and, thus, the primary analytic approach for our real-data application. The corresponding statistical analysis plans for the running examples are given in the eAppendices; <http://links.lww.com/EDE/C169> and include the design and results of the simulation study. As illustrated in Table 1, the interpretation of the study results must account for the statistical assumptions of the selected estimator as well as the plausibility of the identifiability assumptions. Altogether, our approach facilitates objective selection and implementation of the best analysis to answer our research question, while protecting research integrity by ensuring we remain blinded to the true causal effect. Importantly, our approach is in line with regulatory guidelines to update the statistical analysis plan based on a blinded review of the data.⁶

Our proposed process has several consequences for improving research transparency and reproducibility. First, the simulation leads to a fully prespecified statistical analysis plan, where analytic decisions are clearly stated and can be critically evaluated. Second, conducting the simulation requires implementing all candidate estimators in computing code. Thus, these simulations serve as an invaluable tool for debugging code and identifying potential issues (e.g., lack of convergence due to rare outcomes) that may arise in the real-data application. Uploading the computing code and results from both the simulation study and real-data analysis to an online repository, such as GitHub, and including detailed explanations through a markup language further improve reproducibility, trust, and open science.

DISCUSSION

For real-data analyses, we have outlined how simulations can guide the objective selection of the optimal approach for estimation and inference and, thus, full prespecification of the statistical analysis plan. Anchored on the Roadmap, these simulations are designed with our research question at the forefront and to explore the primary concerns of the real-data application. The results of the simulation may ultimately reveal that it is not feasible to reliably estimate the statistical

estimand of interest. In such cases, we may need to return to the early steps of the Roadmap and modify the research question and causal estimand to accommodate the limitations in the real data.⁴⁵

Further guidance is needed, and our presentation has limitations. First, we focused on plasmode simulations and did not cover alternative approaches, which might be needed if the real data are not available or only partially available. Additionally, we emphasized the need to emulate the real data closely but did not discuss how to assess the quality of the emulation. It is worth noting that creating simulated data that are “too close” to the true distribution can inspire fears of data “snooping”;⁶⁹ prespecification and code sharing can help alleviate these fears. Third, our presentation did not cover practical implementation, such as how to vary simulation parameters, determine the number of iterations, and parallelize; we refer to Morris et al.²³ for an excellent overview of these and other considerations. Additional practical details on the projects inspiring our running examples are available in Nance et al.⁴³ and Balzer et al.⁷⁰ Fourth, while our worked examples incorporated common challenges, including confounding, dependence, and missing data, they did not cover other concerns such as generalizability, transportability, and partial identification in detail.^{71–75} Finally, we have presented a two-step process for estimator selection and implementation: (1) conduct a realistic simulation study to objectively compare prespecified estimators according to prespecified metrics and (2) implement the optimal estimator (as defined by the simulation study) for the real-data analysis. Alternative approaches, such as auto-TMLE, are being developed to dynamically evaluate estimators in simulations and implement the optimal estimator in a single step.^{76,77}

Altogether, we believe simulations, anchored on the Roadmap, are an indispensable and underutilized tool for the objective comparison of approaches for estimation and inference in real-data applications. They are a crucial alternative to the status quo: naively applying a preferred algorithm or trying several algorithms and selecting the “best” in an ad hoc manner. Instead, our approach provides a formal framework for comparative assessment of alternative strategies for estimation and inference, prespecification of the corresponding statistical analysis plan, and generating quality-controlled computing code. Our approach strives to improve the transparency, rigor, and reproducibility of real-data analyses in epidemiology and beyond.

ACKNOWLEDGMENTS

We gratefully acknowledge the JICI collaborators who inspired the diabetes data example as well as the SEARCH-Youth study team and participants who inspired the randomized trial example. We also thank Edward Bein and Andrew Mertens for their generous feedback on the manuscript draft.

REFERENCES

- Petersen ML, van der Laan MJ. Causal models and learning from data. *Epidemiology*. 2014;25:418–426.
- van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media; 2011.
- Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183:758–764.
- Fox MP, Edwards JK, Platt R, Balzer LB. The critical importance of asking good questions: the role of epidemiology doctoral training programs. *Am J Epidemiol*. 2020;189:261–264.
- Dang LE, Balzer LB. Start with the target trial protocol; then follow the roadmap for causal inference. *Epidemiology*. 2023;34:619–623.
- ICH Harmonised Tripartite Guideline. *Statistical Principles for Clinical Trials E9*. 1998. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf.
- European Medicines Agency. *ICH E9 (R1) addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials*. 2020. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf.
- U.S. Food and Drug Administration (FDA). *Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry. Technical report*. 2023. Available at: <https://www.fda.gov/media/148910/download>. Accessed 7 November 2023.
- National Institutes of Health. NOT-OD-16-149: NIH Policy on the Dissemination of NIH-Funded Clinical Trial Information. 2017. Accessed 7 November 2023.
- Dang LE, Gruber S, Lee H, et al. A Causal Roadmap for generating high-quality real-world evidence. *J Clin Transl Sci*. 2023;7:e212.
- Munafò MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. *Nat Hum Behav*. 2017;1.
- Hiemstra B, Keus F, Wetterslev J, Gluud C, van der Horst ICC. DEBATE-statistical analysis plans for observational studies. *BMC Med Res Methodol*. 2019;19:233.
- Díaz I, Lee H, Kiciman E, et al. Sensitivity analysis for causality in observational studies for regulatory science. *J Clin Transl Sci*. 2023;7:e267.
- Mathur MB, Fox MP. Toward open and reproducible epidemiology. *Am J Epidemiol*. 2023;192:658–664.
- Gruber S, Lee H, Phillips R, Ho M, van der Laan M. Developing a targeted learning-based statistical analysis plan. *Stat Biopharm Res*. 2023;15:468–475.
- Dang LE, Fong E, Tarp JM, et al. Case study of semaglutide and cardiovascular outcomes: An application of the Causal Roadmap to a hybrid design for augmenting an RCT control arm with real-world data. *J Clin Transl Sci*. 2023;7:e231.
- Ruel T, Mwangwa F, Balzer LB, et al. A multilevel health system intervention for virological suppression in adolescents and young adults living with HIV in rural Kenya and Uganda (SEARCH-Youth): a cluster randomized trial. *Lancet HIV*. 2023;10:e518–e527.
- 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. *Math Model*. 1986;7:1393–1512.
- Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61:962–973.
- Balzer LB, van der Laan M, Ayieko J, et al. Two-stage TMLE to reduce bias and improve efficiency in cluster randomized trials. *Biostatistics*. 2021;24:502–517.
- van der Laan MJ, Rose S. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer; 2018.
- Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan M. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *J Causal Inference*. 2014;2:147–185.
- Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38:2074–2102.
- Tsiatis AA, Davidian M. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22:523–539.
- Sekhon JS, Gruber S, Porter KE, van der Laan MJ. Propensity-score-based estimators and C-TMLE. In van der Laan MJ, Rose S, eds, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011: 343–364.
- Naimi AI, Mishler AE, Kennedy EH. Challenges in obtaining valid causal effect estimates with machine learning algorithms. *Am J Epidemiol*. 2021;192:1536–1544.
- Balzer LB, Westling T. Demystifying statistical inference when using machine learning in causal research. *Am J Epidemiol*. 2021;192:1545–1549.
- Dorie V, Hill J, Shalit U, Scott M, Cervone D. Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Stat Sci*. 2019;34:43–68.
- Fox MP, Nianogo R, Rudolph JE, Howe CJ. Illustrating how to simulate data from directed acyclic graphs to understand epidemiologic concepts. *Am J Epidemiol*. 2022;191:1300–1306.
- Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS One*. 2020;15:e0230405.
- Althoff KN, Chandran A, Zhang J, et al; North American AIDS Cohort Collaboration on Research and Design (NA-ACCORD) of IeDEA. Life-expectancy disparities among adults with HIV in the United States and Canada: the impact of a reduction in drug- and alcohol-related deaths using the lives saved simulation model. *Am J Epidemiol*. 2019;188:2097–2109.
- Nianogo RA, Arah OA. Investigating the role of childhood adiposity in the development of adult type 2 diabetes in a 64-year follow-up cohort: an application of the parametric G-formula within an agent-based simulation study. *Epidemiology*. 2019;30:S101–S109.
- Bykov K, Franklin JM, Li H, Gagne JJ. Comparison of self-controlled designs for evaluating outcomes of drug-drug interactions: simulation study. *Epidemiology*. 2019;30:861–866.
- Zivich PN, Edwards JK, Lofgren ET, Cole SR, Shook-Sa BE, Lessler J. Transportability without positivity: a synthesis of statistical and simulation modeling. *Epidemiology*. 2024;35:23–31.
- Filshstein TJ, Li X, Zimmerman SC, Ackley SF, Glymour MM, Power MC. Proof of concept example for use of simulation to allow data pooling despite privacy restrictions. *Epidemiology*. 2021;32:638–647.
- Chow S-C, Shao J, Wang H, Lokhnygina Y. *Sample Size Calculations in Clinical Research*. 3rd ed. Chapman and Hall/CRC; 2017.
- Office of the Commissioner. *Modeling & Simulation at FDA*. FDA; 2022.
- Balzer LB, Havlir DV, Schwab J, Van Der Laan MJ, Petersen ML. Statistical analysis plan for SEARCH Phase I: health outcomes among adults. *arXiv*. 2018. arXiv:1808.03231.
- Fox MP, MacLehose RF, Lash TL. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer; 2021.
- Jayaweera RT, Bradshaw PT, Gerds C, et al. Accounting for misclassification and selection bias in estimating effectiveness of self-managed medication abortion. *Epidemiology*. 2023;34:140–149.
- Montoya LM, Kosorok MR, Geng EH, Schwab J, Odeny TA, Petersen ML. Efficient and robust approaches for analysis of sequential multiple assignment randomized trials: illustration using the ADAPT-R trial. *Biometrics*. 2023;79:2577–2591.
- Williamson BD, Wyss R, Stuart EA, et al. An application of the Causal Roadmap in two safety monitoring case studies: causal inference and outcome prediction using electronic health record data. *J Clin Transl Sci*. 2023;7:e208.
- Nance N, Mertens A, Gerds T, et al. Applying the Causal Roadmap to longitudinal national Danish registry data: a case study of second-line diabetes medication and dementia. *arXiv*. 2023. arXiv:2310.03235.
- Boulesteix AL, Groenwold RH, Abrahamowicz M, et al; STRATOS Simulation Panel. Introduction to statistical simulations in health research. *BMJ Open*. 2020;10:e039921.
- Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*. 2012;21:31–54.
- Rudolph KE, Gimbrone C, Matthey EC, et al. When effects cannot be estimated: redefining estimands to understand the effects of naloxone access laws. *Epidemiology*. 2022;33:689–698.
- Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*. 1952;47:663–685.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–55.

49. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc.* 1994;89:846–866.
50. Benitez A, Petersen ML, van der Laan MJ, et al. Defining and estimating effects in cluster randomized trials: a methods comparison. *Stat Med.* 2023;42:3443–3466.
51. Nugent JR, Marquez C, Charlebois ED, Abbott R, Balzer LB. Blurring cluster randomized trials and observational studies: two-stage TMLE for subsampling, missingness, and few independent units. *Biostatistics.* 2024;25:599–616.
52. Kahan BC, Forbes G, Ali Y, et al. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials.* 2016;17:438.
53. Hubbard AE, Ahern J, Fleischer NL, et al. To GEE or not to GEE comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology.* 2010;21:467–474.
54. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis*. Chapman and Hall/CRC; 2008:567–614.
55. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol.* 2007;6. doi:10.2202/1544-6115.1309.
56. Phillips RV, van der Laan MJ, Lee H, Gruber S. Practical considerations for specifying a super learner. *Int J Epidemiol.* 2023;52:1276–1285.
57. Gruber S, Phillips RV, Lee H, van der Laan MJ. Data-adaptive selection of the propensity score truncation level for inverse-probability-weighted and targeted maximum likelihood estimators of marginal point treatment effects. *Am J Epidemiol.* 2022;191:1640–1651.
58. Balzer LB, van der Laan MJ, Petersen ML; SEARCH Collaboration. Adaptive pre-specification in randomized trials with and without pair-matching. *Stat Med.* 2016;35:4528–4545.
59. Benkeser D, Carone M, Laan MJVD, Gilbert PB. Doubly robust nonparametric inference on the average treatment effect. *Biometrika.* 2017;104:863–880.
60. Tran L, Petersen M, Schwab J, van der Laan MJ. Robust variance estimation and inference for causal effect estimation. *J Causal Inference.* 2023;11. doi:10.1515/jci-2021-0067.
61. Balzer LB, Cai E, Godoy Garraza L, Amaranath P. Adaptive selection of the optimal strategy to improve precision and power in randomized trials. *Biometrics.* 2024;80:ujad034.
62. Lendle SM, Schwab J, Petersen ML, van der Laan MJ. Itmle: an R package implementing targeted minimum loss-based estimation for longitudinal data. *J Stat Softw.* 2017;81:1–21.
63. Rubinstejn RY. *Simulation and the Monte Carlo method*. 3rd ed. Wiley series in probability and statistics; 2017.
64. Schreck N, Slynko A, Saadati M, Benner A. Statistical plasmode simulations – potentials, challenges and recommendations. *Stat Med.* 2024;43:1804–1825.
65. Benkeser D, Diaz I, Luedtke A, Segal J, Scharfstein D, Rosenblum M. Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics.* 2021;77:1467–1481.
66. Wyss R, Schneeweiss S, Lin KJ, Miller DP, Kalilani L, Franklin JM. Synthetic negative controls: using simulation to screen large-scale propensity score analyses. *Epidemiology.* 2022;33:541–550.
67. Benkeser D, Van Der Laan M. The highly adaptive lasso estimator. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2016:689–696.
68. van der Vaart AW. *Asymptotic Statistics*. Cambridge University Press; 2000.
69. Fisher A. *Treatment Effect Bias From Sample Snooping: Blinding Outcomes is Neither Necessary nor Sufficient*. 2020. Available at: <https://arxiv.org/abs/2007.02514>.
70. Balzer LB, Ruel T, Havlir DV; the SEARCH-Youth Study Team. Statistical analysis plan for primary and selected secondary health endpoints of the SEARCH-Youth study. *arXiv.* 2022. Available at: <https://arxiv.org/abs/2211.02771>.
71. Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: *Health Service Research Methodology: A Focus on AIDS*. US Public Health Service; 1989:113–159.
72. Manski CF. Nonparametric bounds on treatment effects. *Am Econ Rev.* 1990;80:319–323.
73. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *J R Stat Soc Ser A Stat Soc.* 2011;174:369–386.
74. Bareinboim E, Pearl J. A general algorithm for deciding transportability of experimental results. *J Causal Inference.* 2013;1:107–134.
75. Swanson SA, Hernán MA, Miller M, Robins JM, Richardson TS. Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes. *J Am Stat Assoc.* 2018;113:933–947.
76. Shortreed SM, Moodie EEM. Automated analyses: because we can, does it mean we should? *Stat Sci.* 2020;35:499–502.
77. Benkeser D, Cai W, van der Laan MJ. A nonparametric super-efficient estimator of the average treatment effect. *Stat Sci.* 2020;35:484–495.