This is a "preproof" accepted article for *Journal of Clinical and Translational Science*.
This version may be subject to change during the production process.
10.1017/cts.2023.635

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

Lauren E Dang[1], Susan Gruber[2], Hana Lee[3], Issa J Dahabreh[4], Elizabeth A Stuart[5], Brian D Williamson[6], Richard Wyss[7], Iván Díaz[8], Debashis Ghosh[9], Emre Kıcıman[10], Demissie Alemayehu[11], Katherine L Hoffman[12], Carla Y Vossen[13], Raymond A Huml[14], Henrik Ravn[15], Kajsa Kvist[15], Richard Pratley[16], Mei-Chiung Shih[17,18], Gene Pennello[19], David Martin[20], Salina P Waddy[21], Charles E Barr[22,23], Mouna Akacha[24], John B Buse[25], Mark van der Laan*[1], Maya Petersen*[1]

[1]Department of Biostatistics, University of California, Berkeley, CA, USA
[2]TL Revolution, Cambridge, MA, USA
[3]Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA
[4]CAUSALab, Department of Epidemiology and Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[5]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[6]Biostatistics Division, Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA
[7]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[8]Division of Biostatistics, Department of Population Health, New York University Grossman School of Medicine, New York, NY, USA

[9]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
[10]Microsoft Research, Redmond, WA, USA
[11]Global Biometrics and Data Management, Pfizer Inc., New York, NY, USA
[12]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA
[13]Syneos Health Clinical Solutions, Amsterdam, The Netherlands
[14]Syneos Health Clinical Solutions, Morrisville, NC, USA
[15]Novo Nordisk, Søborg, Denmark
[16]AdventHealth Translational Research Institute, Orlando, FL, USA
[17]Cooperative Studies Program Coordinating Center, VA Palo Alto Health Care System, Palo Alto, CA, USA
[18]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
[19]Division of Imaging Diagnostics and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD, USA
[20]Global Real World Evidence Group, Moderna, Cambridge, MA, USA
[21]National Center for Advancing Translational Sciences, Bethesda, MD, USA
[22]Graticule Inc., Newton, MA, USA
[23]Adaptic Health Inc., Palo Alto, CA, USA
[24]Novartis Pharma AG, Basel, Switzerland
[25]Division of Endocrinology, Department of Medicine, University of North Carolina, Chapel Hill, NC, USA

*****Co-Senior Authors

Corresponding Author: Lauren Eyler Dang, Division of Biostatistics, University of California, Berkeley, 2121 Berkeley Way, Room 5302, Berkeley, CA 94720, (510) 642-3241, Lauren.eyler@berkeley.edu

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Conflict of Interest Statement:

LED reports tuition and stipend support from a philanthropic gift from the Novo Nordisk corporation to the University of California, Berkeley to support the Joint Initiative for Causal Inference. **IJD** is the principal investigator of a research agreement between Harvard University and Sanofi and reports consulting fees from Moderna. **EK** is employed by Microsoft. **DA** is employed by Pfizer Inc. and holds stocks in Pfizer Inc. **CYV** and **RAH** are employed by Syneos Health. **CYV** reports that her husband is employed by Galapagos. **HR** and **KK** are employed by Novo Nordisk A/S and own stocks in Novo Nordisk A/S. **RP** has received the following (directed to his institution): speaker fees from Merck and Novo Nordisk; consulting fees from Bayer AG, Corcept Therapeutics Incorporated, Dexcom, Endogenex, Inc., Gasherbrum Bio, Inc., Hanmi Pharmaceutical Co., Hengrui (USA) Ltd., Lilly, Merck, Novo Nordisk, Pfizer, Rivus Pharmaceuticals Inc., Sanofi, Scohia Pharma Inc., and Sun Pharmaceutical Industries; and grants from Hanmi Pharmaceuticals Co., Metavention, Novo Nordisk, and Poxel SA. **DM** is employed by Moderna. **CEB** is co-founder of Adaptic Health Inc., Managing Director of Pivotal Strategic Consulting, LLC, and receives consulting fees from Graticule Inc. and Sophic Alliance Inc. **MA** is employed by Novartis Pharma AG. **JBB** reports contracted fees and travel support for contracted activities for consulting work paid to the University of North Carolina by Novo Nordisk; grant support by Dexcom, NovaTarg, Novo Nordisk, Sanofi, Tolerion and vTv Therapeutics; personal compensation for consultation from Alkahest, Altimmune, Anji, AstraZeneca, Bayer, Biomea Fusion Inc, Boehringer-Ingelheim, CeQur, Cirius Therapeutics Inc, Corcept Therapeutics, Eli Lilly, Fortress Biotech, GentiBio, Glycadia, Glyscend, Janssen, MannKind, Mellitus Health, Moderna, Pendulum Therapeutics, Praetego, Sanofi, Stability Health, Terns Inc, Valo and Zealand Pharma; stock/options in Glyscend, Mellitus Health, Pendulum Therapeutics, PhaseBio, Praetego, and Stability Health; and board membership of the Association of Clinical and Translational Science. **MvdL and SG** report that they are co-founders of the statistical software start-up company TLrevolution, Inc. **MvdL and MP** report personal compensation for consultation from Novo Nordisk.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Abstract

Increasing emphasis on the use of real-world evidence (RWE) to support clinical policy and regulatory decision-making has led to a proliferation of guidance, advice, and frameworks from regulatory agencies, academia, professional societies, and industry. A broad spectrum of studies use real-world data (RWD) to produce RWE, ranging from randomized trials with outcomes assessed using RWD to fully observational studies. Yet, many proposals for generating RWE lack sufficient detail, and many analyses of RWD suffer from implausible assumptions, other methodological flaws, or inappropriate interpretations. The *Causal Roadmap* is an explicit, itemized, iterative process that guides investigators to pre-specify study design and analysis plans; it addresses a wide range of guidance within a single framework. By supporting the transparent evaluation of causal assumptions and facilitating objective comparisons of design and analysis choices based on pre-specified criteria, the *Roadmap* can help investigators to evaluate the quality of evidence that a given study is likely to produce, specify a study to generate high-quality RWE, and communicate effectively with regulatory agencies and other stakeholders. This paper aims to disseminate and extend the *Causal Roadmap* framework for use by clinical and translational researchers; three companion papers demonstrate applications of the *Causal Roadmap* for specific use cases.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Introduction

The 21st century has witnessed a dramatic increase in the quality, diversity, and availability of real-world data (RWD) such as electronic health records, health insurance claims, and registry data [1]. 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 [2] 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" [3]. Internationally, stakeholders including other regulatory agencies, industry, payers, academia, and patient groups have also increasingly endorsed the use of RWE to support regulatory decisions [4,5]. Study designs that use RWD to generate RWE (referred to below as RWE studies) include pragmatic clinical trials, externally controlled trials or hybrid randomized-external data studies, and fully observational studies [6–8].

There are multiple motivations for using RWD in a study. First, RWD 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 for reasons including strict eligibility criteria, strict treatment protocols, limited patient numbers, and limited time on treatment and in follow-up [9]. Second, recent drug development efforts have more commonly targeted rare diseases or conditions without effective treatments [10]. RWD can be useful in such contexts when it is not practical to randomize enough participants to power a standard randomized trial or when there is an ethical imperative to minimize the number of patients assigned to the trial control arm [11,12]. RWD was also highly valuable during the COVID-19 pandemic; observational studies reported timely evidence on vaccine booster effectiveness [13,14], the comparative effectiveness of different vaccines [15], and vaccine effectiveness during pregnancy [16].

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

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

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 proliferation of guidance documents and framework proposals from regulatory agencies, academia, and industry addressing different stages of the process of RWE generation [3,5,17–23]. 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 [20]. To address this gap between guidance and implementation and to discuss perspectives from regulatory and federal medical research agencies, industry, academia, trialists, methodologists, and software developers, the Forum on the Integration of Observational and Randomized Data (FIORD) meeting was held in Washington, D.C. November 17-18, 2022. FIORD participants discussed their experiences with RWE guidance and best practices, as well as steps that could be taken to help investigators follow available guidance. Specifically, participants determined the need for a unifying structure to assist with specification of key elements of a design and analysis plan 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* [24–30] (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 key steps necessary for pre-specifying RWE study design and analysis plans 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) specification of key elements of a study design and analysis plan that is expected 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 credible RWE in the future.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

The goal of this paper is to disseminate the *Causal Roadmap* 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. Members of the FIORD Working Groups also provide three case studies as companion papers demonstrating application of the *Roadmap*, as described in Table 1.

Overview of the *Causal Roadmap* for clinical and translational scientists

We walk through the steps of the *Roadmap*, depicted in Figure 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 *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 [24,31–36]. The current paper is not a comprehensive introduction, but rather aims to describe a structured approach that can support the generation of high-quality evidence.

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

The first step involves defining the causal question, causal model, and the causal estimand that would answer the question. To facilitate explanation of these concepts, we start by using frameworks for specifying components of a causal question and estimand to also specify key elements of the causal model (Step 1a) before further elaborating the causal model in Step 1b.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

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 [36–40], both the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9(R1) [41] and Target Trial Emulation [17,32,42,43] frameworks prompt investigators to define components of a causal question and estimand. The causal estimand is a mathematical quantity that represents the answer to the causal question (Table 2).

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 received the drug under investigation (treatment strategy $A = 1$) versus an active comparator (treatment strategy $A = 0$) 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 [44,45] individuals would have experienced had they received treatment strategy $A = 1$ and not been censored ($Y^{a=1,c=0}$) *and* the potential outcomes the same individuals would have experienced had they received treatment strategy $A = 0$ and not been censored ($Y^{a=0,c=0}$).

A more fully elaborated structural causal model would help us describe the causal pathways that generate these potential outcomes [46]. 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"),
$\Psi^* = P(Y^{a=1,c=0} = 1) - P(Y^{a=0,c=0} = 1)$.

This mathematical quantity – a function of the potential outcomes defined above – is the *causal estimand* of interest in our example. Table 2 lists other examples of causal estimands.

Importance: Even though we can only observe at most one potential outcome for each individual [47], 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

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

in Table 2 is crucial for specifying a study design and analysis plan to provide the best possible effect estimate. Ultimately, we need a procedure that can be applied to the data to generate an appropriate estimate (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: Experts in causal inference may come from a variety of fields, including economics, biostatistics, epidemiology, computational science, the social sciences, medicine, pharmacology, and others. They can help to translate a research question into a causal estimand.

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

Next, we consider what we know (and do not know) about the processes that will generate – or that have already generated – data to answer this question. First, we consider the type of study (e.g., pragmatic randomized trial, retrospective cohort study). Then, we consider what factors affect the variables that are part of our treatment strategies – found in Table 2 and referred to as intervention variables below – and the outcome in our proposed study. It is also important to consider factors that are affected by intervention and/or outcome variables, such as mediators, colliders, or any study eligibility criteria that are outcome-dependent.

This background knowledge helps to generate the causal model [46]. We specified some key variables in our causal model in Step 1a (in Table 2 and our potential outcomes). Now, we further elaborate our causal model by describing potential causal relationships between these and other important variables. Multiple tools and frameworks can help elicit this information, such as conceptual models and causal graphs (e.g., directed acyclic graphs or single world intervention graphs) [39,48–51].

Figure 2 gives a simple example of causal graph construction for a prospective observational cohort study, 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

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

$Y^{\star}$, which is equal to the actual outcome if it was observed and is missing otherwise (Figure 2a). Arrows denote possible effects of one variable on another.

Then, we attempt to write down factors that might influence (or be influenced by) these variables. Figure 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, participant characteristics or other non-randomized factors (such as policy or environmental factors) may affect both the treatment/exposure and the outcome. Next, we consider factors that are unmeasured or difficult to measure that might influence treatment, outcomes, or censoring. Figure 2c shows access to healthcare as an example.

Causal graphs can become much more complicated, especially when working with longitudinal data [52], using proxies for unmeasured variables [53], or combining different data sources [54] (as demonstrated in the case study of Semaglutide and Cardiovascular Outcomes). A carefully constructed causal graph should also include sample selection, competing risks, intercurrent events, and measurement error [32,55]. Examples of causal graph construction are available for a wide variety of study designs including retrospective cohort, cross-sectional, and case-control studies in which selection into the study sample may be affected by the outcome [56,57].

Importance: Considering which factors may affect or be affected by 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 [32].

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.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Consider whether the causal question and estimand (Step 1a) need to be modified based on Step 1b.

After writing down our causal model, we sometimes need to change our question [58]. 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 outcome or death [41]. ICH E9(R1) discusses other intercurrent events and alternative estimands [41].

Step 2: Consider 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 consider 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 [42]?

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 tertiary facilities* 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 [3] and whether we are able to estimate a causal effect from the observed data (discussed in Step 3).

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Build a Multidisciplinary Collaboration: Clinicians and clinical informaticists can help to explain the way variables are measured in relation to underlying medical concepts or in relation to a particular care setting. Statisticians can help to determine how to match baseline time zero in the observed data with the follow-up period in the causal question.

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

In Step 3, we ask whether the data we 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 express the causal estimand as a function of the observed data distribution (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 [27]. If there is a causal gap, even the true value 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 randomized trials, 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 lists two examples of identification assumptions with informal explanations of their meaning.

Exchangeability, in particular, can also be framed in terms of causal graphs [46]. Confounding by unmeasured variables is a widely discussed source of bias in observational studies. Conditioning on a variable that is independently affected by both treatment and the outcome – either by adjusting for that variable in the analysis or by selecting retrospective study participants based on certain values of that variable – may also result in a non-causal association between

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

treatment and outcomes and a statistical estimand that is biased for the true causal effect (i.e., collider bias or selection bias) [32].

Depending on the causal model and question (Step 1), additional assumptions or alternate sets of assumptions may be necessary. For example, if we aim to transport or generalize a causal effect to a new population, we must assume that all values of effect modifiers represented in the target population are also represented in the original study population and that all effect modifiers with different distributions in these two populations are measured [59–63]. Hernán and Robins (2020) [32], among others, provide in-depth discussions of identification assumptions. The three companion papers demonstrate the evaluation of these assumptions.

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 [64,65]. 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 [66]. If multiple study designs are feasible, Step 3 can help us to consider which study design is based on more reasonable assumptions [67].

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 [17,32,37]. 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 [68,69]. 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.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

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 randomized trials that do not utilize RWD due to issues such as informative missingness [32]. Nevertheless, 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: Experts in causal inference can aid other investigators in evaluating different causal identification assumptions. For example, reasoning about the exchangeability assumption can become quite complicated if there are multiple intervention variables (e.g., when the treatment varies over time) [39,52]. 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 [39,46,70]. Software programs can also facilitate this process [71,72].

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:

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

In a simple case where participant characteristics other than our intervention variables and outcome – 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

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

In words, we have re-written the answer to our causal question (which is defined based on potential outcomes that we cannot simultaneously observe) in terms of a quantity that we can

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

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 correctly specified (a questionable assumption, as discussed below) [24]. 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 [73]. 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 [74–76]. Instead, the *Roadmap* involves choosing a statistical estimand that corresponds to the causal estimand under identification assumptions. 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 [39]. 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 [39].

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-

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

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 [24].

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 [24]. Examples include type I 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 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 [24].

Many commonly used 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 [24,73,77]. The traditional practice of defaulting to a parametric regression as the statistical estimator imposes additional 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., based on 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) [24].

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 [24]. This bias may result in misleading conclusions. We aim to choose an estimator that not only has minimal bias but also is

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

efficient – thereby producing 95% confidence intervals that are accurate but as narrow as possible – to make maximal use of the data [24].

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 [78]. We use the term "outcome-blind" to mean that the simulations are conducted without information on the observed treatment-outcome association in the current study; such simulations may utilize other information from previously collected data or from the current study data if available (e.g., data on baseline covariates, treatment, and censoring) to approximate the real experiment [78]. Simulations conducted before data collection may use a range of plausible values for these study characteristics [79]. As recommended by ICH E9(R1), simulations should also be conducted for cases involving plausible violations of the statistical assumptions underpinning the estimators [41]. Examples of such violations include non-linearity for linear models or inaccurate prior distributions for Bayesian parameters. For an example of conducting such a simulation, please see the Drug Safety and Monitoring case study.

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) [80] provide a detailed description of how to pre-specify a SAP using targeted minimum loss-based estimation (TMLE) [81] and super learning [77], a combined approach that integrates machine learning to minimize the chance that statistical modeling assumptions are violated [24].

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 [32,68,82–84]. 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 [41]. One

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

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 [68,83–85]. Sensitivity analysis also allows for construction of confidence intervals that account for plausible values of the causal gap [27,68,83–85]. Alternatively, investigators may assess for causal bias using negative control variables [86-87].

The specifics of these methods – and alternative approaches – are beyond the scope of this paper, but the case study of Nifurtimox for Chagas Disease in the companion paper provides an overview of methods for sensitivity analysis, as well as a worked example of using available evidence to assess a plausible range for the causal gap. As discussed in this case study, the method of sensitivity analysis should be pre-specified prior to estimating the effect of interest [88]. 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 [83].

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 [27,68,69,89]. 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 [69,90]; this was the case in Cornfield et al. (1959)'s seminal sensitivity analysis of the effect of smoking on lung cancer [91]. 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 [69,92].

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 [83]**.**

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Step 7: Compare alternative study designs

Roadmap Steps 1-6 help us to specify a study design and analysis plan, 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 design could fall anywhere on the spectrum from a traditional randomized trial to a fully observational analysis. In cases when it is not possible to conduct a traditional randomized trial due to logistical or ethical reasons – or when trial 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 [32]. If conducting a randomized trial is feasible, baseline randomization of an intervention (as part of either a traditional or pragmatic trial [93]) still generally affords a higher degree of certainty that the estimated effect has a causal interpretation 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 as opposed to a traditional randomized trial 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 [94–96]. 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 I error and power, but also metrics quantifying how the proposed designs will modify the medical product development process [94]. The case study of Semaglutide and Cardiovascular Outcomes demonstrates how to compare study designs that are based on *Roadmap* Steps 1-6.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

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 a randomized trial with external data. When comparing the standard and augmented randomized trial designs, 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 I error that would be expected if inappropriate external controls are considered [96]. 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 [67]. Simulated quantification of these tradeoffs using pre-specified benchmarks can help investigators to make design choices transparent [97].

Build a Multidisciplinary Collaboration: Factors to consider when comparing different designs include the expected magnitude of benefit based on prior data and the quality of that data [11], the plausible bounds on the causal gap for a given RWE study, the treatments that are currently available [11], and preferences regarding tradeoffs between design characteristics such as type I versus type II error control [97]. Because these tradeoffs will be context-dependent [11,97], collaboration with patient groups and discussion with regulatory agencies is often valuable when choosing a study design from multiple potential options.

A list of *Roadmap* steps for specifying key elements of a study design and analysis plan

Table 4 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.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Discussion

The *Roadmap* can help investigators to pre-specify design and analysis plans 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 companion papers presenting case studies on Drug Safety and Monitoring, Nifurtimox for Chagas Disease, and Semaglutide and Cardiovascular Outcomes demonstrate applications of the *Roadmap* and explain specific steps in greater detail.

Past descriptions of the *Roadmap* have largely been targeted to quantitative scientists [24–27,29,30]. In this paper, we focus on intuitive explanations rather than formal mathematical results to make these causal inference concepts more accessible to a wide audience. We 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 [98] and emphasizes the quantitative comparison of the proposed study 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 regulatory guidance documents, as well; these include the FDA's Framework and Draft Guidance documents for RWE that emphasize the quality and appropriateness of the data [3,99–101] and the ICH E9(R1) guidance on estimands and sensitivity analysis [41]. 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 [44,45], the careful design of non-experimental studies [35,36,38,40], causal graphs [39,48–51] and structural causal models [46], causal identification

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

[39,46,102], translation of causal to statistical estimands using the g-formula [39], and methods for estimation and sensitivity analysis [24,34,68,77,82,84]. The *Roadmap* is also compatible with frameworks including the Target Trial Emulation framework [17,43], the Patient-Centered Outcomes Research Institute (PCORI) Methodology Standards [19], white papers from the Duke-Margolis Center [18,103], the REporting of studies Conducted using Observational Routinely-Collected health Data (RECORD) Statement [104], the Structured Preapproval and Postapproval Comparative study design framework [105], and the STaRT-RWE template [20]. 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 key 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 randomized trial to a fully observational analysis.

With increasing emphasis by regulatory agencies around the world regarding the importance of RWE [5], 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" [106]. By following the *Roadmap* steps to pre-specify a study design and analysis plan, investigators can set themselves up to convey relevant information to regulators and other stakeholders, to produce high-quality estimates of causal effects using RWD when possible, and to honestly evaluate whether the proposed methods are adequate for drawing causal inferences.

Disclaimer: The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS, the U.S. Government, or the authors' affiliations.

Acknowledgements: We would like to thank the sponsors of the FIORD workshop, including the Forum for Collaborative Research and the Center for Targeted Machine Learning and Causal Inference (both at the School of Public Health at the University of California, Berkeley), and the Joint Initiative for Causal Inference. We would also like to thank Dr. John Concato (U.S. Food

and Drug Administration) for his comments on this manuscript. This research was funded by a philanthropic gift from the Novo Nordisk corporation to the University of California, Berkeley to support the Joint Initiative for Causal Inference.

Disclosures: LED reports tuition and stipend support from a philanthropic gift from the Novo Nordisk corporation to the University of California, Berkeley to support the Joint Initiative for Causal Inference. **IJD** is the principal investigator of a research agreement between Harvard University and Sanofi and reports consulting fees from Moderna. **EK** is employed by Microsoft. **DA** is employed by Pfizer Inc. and holds stocks in Pfizer Inc. **CYV** and **RAH** are employed by Syneos Health. **CYV** reports that her husband is employed by Galapagos. **HR** and **KK** are employed by Novo Nordisk A/S and own stocks in Novo Nordisk A/S. **RP** has received the following (directed to his institution): speaker fees from Merck and Novo Nordisk; consulting fees from Bayer AG, Corcept Therapeutics Incorporated, Dexcom, Endogenex, Inc., Gasherbrum Bio, Inc., Hanmi Pharmaceutical Co., Hengrui (USA) Ltd., Lilly, Merck, Novo Nordisk, Pfizer, Rivus Pharmaceuticals Inc., Sanofi, Scohia Pharma Inc., and Sun Pharmaceutical Industries; and grants from Hanmi Pharmaceuticals Co., Metavention, Novo Nordisk, and Poxel SA. **DM** is employed by Moderna. **CEB** is co-founder of Adaptic Health Inc., Managing Director of Pivotal Strategic Consulting, LLC, and receives consulting fees from Graticule Inc. and Sophic Alliance Inc. **MA** is employed by Novartis Pharma AG. **JBB** reports contracted fees and travel support for contracted activities for consulting work paid to the University of North Carolina by Novo Nordisk; grant support by Dexcom, NovaTarg, Novo Nordisk, Sanofi, Tolerion and vTv Therapeutics; personal compensation for consultation from Alkahest, Altimmune, Anji, AstraZeneca, Bayer, Biomea Fusion Inc, Boehringer-Ingelheim, CeQur, Cirius Therapeutics Inc, Corcept Therapeutics, Eli Lilly, Fortress Biotech, GentiBio, Glycadia, Glyscend, Janssen, MannKind, Mellitus Health, Moderna, Pendulum Therapeutics, Praetego, Sanofi, Stability Health, Terns Inc, Valo and Zealand Pharma; stock/options in Glyscend, Mellitus Health, Pendulum Therapeutics, PhaseBio, Praetego, and Stability Health; and board membership of the Association of Clinical and Translational Science. **MvdL and SG** report that they are co-founders of the statistical software start-up company TLrevolution, Inc. **MvdL and MP** report personal compensation for consultation from Novo Nordisk.

References

1. Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. *J Clin Invest*. 2020;130(2):565-574. doi:10.1172/JCI129197

2. 21st Century Cures Act, H.R. 34, 114th Congress. 2016.

3. U.S. Food and Drug Administration. Framework for FDA's real-world evidence program. 2018. https://www.fda.gov/media/120060/download

4. 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 AW, Kahn B, Forstag EH, eds. *Examining the Impact of Real-World Evidence on Medical Product Development: Proceedings of a Workshop Series*. Washington, DC: National Academies Press (US). 2019. https://www.ncbi.nlm.nih.gov/books/NBK540106/

5. Burns L, Roux NL, Kalesnik-Orszulak R, et al. Real-world evidence for regulatory decision-making: Guidance from around the world. *Clin Ther*. 2022;44(3):420-437. doi:10.1016/j.clinthera.2022.01.012

6. Baumfeld Andre E, Reynolds R, Caubel P, Azoulay L, Dreyer NA. Trial designs using real-world data: The changing landscape of the regulatory approval process. *Pharmacoepidemiol Drug Saf*. 2020;29(10):1201-1212. doi:10.1002/pds.4932

7. 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

8. U.S. Food and Drug Administration. *CID case study: External control in diffuse B-cell lymphoma*. 2022. https://www.fda.gov/media/155405/download

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

9. Lavertu A, Vora B, Giacomini KM, Altman R, Rensi S. A new era in pharmacovigilance: Toward real-world data and digital monitoring. *Clin Pharmacol Ther*. 2021;109(5):1197-1202. doi:10.1002/cpt.2172

10. Ringel MS, Scannell JW, Baedeker M, Schulze U. Breaking Eroom's Law. *Nat Rev Drug Discov*. 2020;19(12):833-834. doi:10.1038/d41573-020-00059-3

11. 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

12. Jahanshahi M, Gregg K, Davis G, et al. The use of external controls in FDA regulatory decision making. *Ther Innov Regul Sci*. 2021;55(5):1019-1035. doi:10.1007/s43441-021-00302-y

13. 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. doi:10.1016/S0140-6736(21)02249-2

14. 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. *Lancet Infect Dis*. 2022;22(9):1313-1320. doi:10.1016/S1473-3099(22)00292-4

15. Dickerman BA, Gerlovin H, Madenci AL, et al. Comparative effectiveness of BNT162b2 and mRNA-1273 vaccines in U.S. veterans. *N Engl J Med*. 2022;386(2):105-115. doi:10.1056/NEJMoa2115463

16. Dagan N, Barda N, Biron-Shental T, et al. Effectiveness of the BNT162b2 mRNA COVID-19 vaccine in pregnancy. *Nat Med*. 2021;27(10):1693-1695. doi:10.1038/s41591-021-01490-8

17. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758-764. doi:10.1093/aje/kwv254

18. 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

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

20. Wang SV, 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. doi:10.1136/bmj.m4856

21. Arlett P, Kjær J, Broich K, Cooke E. Real-world evidence in EU medicines regulation: Enabling use and establishing value. *Clin Pharmacol Ther*. 2022;111(1):21-23. doi:10.1002/cpt.2479

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

23. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Epidemiol Camb Mass*. 2007;18(6):800-804. doi:10.1097/EDE.0b013e3181577654

24. van der Laan MJ, Rose S. *Targeted learning: Causal inference for observational and experimental data*. New York, NY: Springer. 2011. doi:10.1007/978-1-4419-9782-1

25. Petersen ML, van der Laan MJ. Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology*. 2014;25(3):418-426. doi:10.1097/EDE.0000000000000078

26. 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. *Stat Biopharm Res*. 2021;15(1):43-56. doi:10.1080/19466315.2021.1883475

27. Gruber S, Phillips RV, Lee H, Ho M, Concato J, van der Laan MJ. Targeted learning: Towards a future informed by real-world evidence. *Stat Biopharm Res*. 2023;00(00):1-15. doi:10.1080/19466315.2023.2182356

28. Petersen ML. Commentary: Applying a causal road map in settings with time-dependent confounding. *Epidemiology*. 2014;25(6):898-901. doi:10.1097/EDE.0000000000000178

29. 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.

30. Saddiki H, Balzer LB. A primer on causality in data science. *J Soc Francaise Stat*. 2020;161(1):67-90.

31. van der Laan MJ, Rose S. *Targeted learning in data science: Causal inference for complex longitudinal studies*. New York, NY: Springer. 2018. doi:10.1007/978-3-319-65304-4

32. Hernan MA, Robins JM. *Causal inference: What if*. Boca Raton, FL: Chapman & Hall/CRC. 2020.

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

34. Lash TL, Fox MP, Fink AK. *Applying quantitative bias analysis to epidemiologic data*. New York, NY: Springer. 2009.

35. Rosenbaum PR. *Observational studies*. New York, NY: Springer. 2nd ed. 2002.

36. Rosenbaum PR. *Design of observational studies*. New York, NY: Springer. 2010.

37. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688-701. doi:10.1037/h0037350

38. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007;26(1):20-36. doi:10.1002/sim.2739

39. 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(9-12):1393-1512. doi:10.1016/0270-0255(86)90088-6

40. Cochran WG, Chambers SP. The planning of observational studies of human populations. *J R Stat Soc Ser Gen*. 1965;128(2):234. doi:10.2307/2344179

41. 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

42. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol*. 2016;79:70-75. doi:10.1016/j.jclinepi.2016.04.014

43. Hernán MA, Wang W, Leaf DE. Target trial emulation: A framework for causal inference from observational data. *JAMA*. 2022;328(24):2446. doi:10.1001/jama.2022.21383

44. 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). *Stat Sci*. 1923;5:465-480.

45. Rubin DB. [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. *Stat Sci*. 1990;5(4). doi:10.1214/ss/1177012032

46. Pearl J. *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press. 2nd ed. 2009. doi:10.1017/CBO9780511803161

47. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945-960. doi:10.1080/01621459.1986.10478354

48. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669-688. doi:10.1093/biomet/82.4.669

49. Wright S. Correlation and causation. *J Agric Res*. 1921;20:557-585.

50. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37-48.

51. 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

52. Robins JM. 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. doi:10.1007/978-1-4612-1842-5_4

53. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. 2019;34(3):211-219. doi:10.1007/s10654-019-00494-6

54. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci*. 2016;113(27):7345-7352. doi:10.1073/pnas.1510507113

55. Stensrud MJ, Dukes O. Translating questions to estimands in randomized clinical trials with intercurrent events. *Stat Med*. 2022;41(16):3211-3228. doi:10.1002/sim.9398

56. Shahar E, Shahar DJ. Causal diagrams and the cross-sectional study. *Clin Epidemiol*. 2013;5:57-65. doi: 10.2147/CLEP.S42843.

57. Mansournia MA, Hernán MA, Greenland S. Matched designs and causal diagrams. *Int J Epidemiol*. 2013;42(3):860-869. doi:10.1093/ije/dyt083

58. Phillips A, Abellan-Andres J, Soren A, et al. Estimands: discussion points from the PSI estimands and sensitivity expert group. *Pharm Stat*. 2017;16(1):6-11. doi:10.1002/pst.1745

59. Degtiar I, Rose S. A review of generalizability and transportability. *Annu Rev Stat Its Appl*. 2023;10(1):501-524. doi:10.1146/annurev-statistics-042522-103837

60. Rudolph KE, Laan MJ. Robust estimation of encouragement design intervention effects transported across sites. *J R Stat Soc Ser B Stat Methodol*. 2017;79(5):1509-1525. doi:10.1111/rssb.12213

61. Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, Hernán MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*. 2019;75(2):685-694. doi:10.1111/biom.13009

62. Dahabreh IJ, Hernán MA. Extending inferences from a randomized trial to a target population. *Eur J Epidemiol*. 2019;34(8):719-722. doi:10.1007/s10654-019-00533-2

63. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. Extending inferences from a randomized trial to a new target population. *Stat Med*. 2020;39(14):1999-2014. doi:10.1002/sim.8426

64. 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(1):31-54. doi:10.1177/0962280210386207

65. Rudolph KE, Gimbrone C, Matthay EC, et al. When effects cannot be estimated: Redefining estimands to understand the effects of naloxone access laws. *Epidemiology*. 2022;33(5):689-698. doi:10.1097/EDE.0000000000001502

66. Gruber S, Phillips RV, 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.

67. Weber AM, van der Laan MJ, Petersen ML. Assumption trade-offs when choosing identification strategies for pre-post treatment effect estimation: An illustration of a community-based intervention in Madagascar. *J Causal Inference*. 2015;3(1):109-130. doi:10.1515/jci-2013-0019

68. Díaz I, van der Laan MJ. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *Int J Biostat*. 2013;9(2). doi:10.1515/ijb-2013-0004

69. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: Introducing the E-Value. *Ann Intern Med*. 2017;167(4):268. doi:10.7326/M16-2607

70. Shpitser I. Complete identification methods for the causal hierarchy. *J Mach Learn Res*. 2008;9:1941-1979.

71. Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GTH. Robust causal inference using directed acyclic graphs: the R package 'dagitty.' *Int J Epidemiol*. 2017;45(6):1887-1894. doi:10.1093/ije/dyw341

72. Sharma A, Kiciman E. DoWhy: An end-to-end library for causal inference. *arXiv*. Preprint posted online November 9, 2020. arXiv:2011.04216.

73. Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat Sci*. 2001;16(3). doi:10.1214/ss/1009213726

74. Hernán MA. The hazards of hazard ratios. *Epidemiology*. 2010;21(1):13-15. doi:10.1097/EDE.0b013e3181c1ea43

75. Greenland S. Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*. 1996;7(5):498-501.

76. Martinussen T, Vansteelandt S, Andersen PK. Subtleties in the interpretation of hazard contrasts. *Lifetime Data Anal*. 2020;26(4):833-855. doi:10.1007/s10985-020-09501-5

77. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol*. 2007;6(1). doi:10.2202/1544-6115.1309

78. 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*. 2022:biom.13808. doi:10.1111/biom.13808

79. 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

80. Gruber S, Lee H, Phillips R, Ho M, van der Laan M. Developing a Targeted Learning-based statistical analysis plan. *Stat Biopharm Res*. 2022;00(00):1-8. doi:10.1080/19466315.2022.2116104

81. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2(1). doi:10.2202/1557-4679.1043

82. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol*. 1996;25(6):1107-1116. doi:10.1093/ije/25.6.1107

83. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol*. 2014;43(6):1969-1985. doi:10.1093/ije/dyu149

84. Robins JM, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME, Berry D, eds. *Statistical models in epidemiology, the environment, and clinical trials*. New York, NY: Springer New York. 2000:1-94. doi:10.1007/978-1-4612-1284-3_1

85. Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J Am Stat Assoc*. 1998;93(444):1321-1339. doi:10.1080/01621459.1998.10473795

86. 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. doi:10.1097/EDE.0b013e3181d61eeb

87. Shi X, Miao W, Tchetgen ET. A selective review of negative control methods in epidemiology. *Curr Epidemiol Rep*. 2020;7(4):190-202. doi:10.1007/s40471-020-00243-4

88. Kasy M, Spiess J. Rationalizing pre-analysis plans: Statistical decisions subject to implementability. *arXiv*. Preprint posted online August 20, 2022. arXiv:2208.09638.

89. Phillips CV, LaPole LM. Quantifying errors without random sampling. *BMC Med Res Methodol*. 2003;3(1):9. doi:10.1186/1471-2288-3-9

90. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J R Stat Soc Ser B Methodol*. 1983;45(2):212-218. doi:10.1111/j.2517-6161.1983.tb01242.x

91. Cornfield J, Haenszel W, Hammond EC, Lilienfeld A, Shimkin M, Wynder E. Smoking and lung cancer: Recent evidence and a discussion of some questions. *JNCI J Natl Cancer Inst*. 1959;22(1):173-203. doi:10.1093/jnci/22.1.173

92. Rudolph KE, Keyes KM. Voluntary firearm divestment and suicide risk: Real-world importance in the absence of causal identification. *Epidemiology*. 2023;34(1):107-110. doi:10.1097/EDE.0000000000001548

93. Ford I, Norrie J. Pragmatic trials. *N Engl J Med*. 2016;375(5):454-463. doi:10.1056/NEJMra1510059

94. Kim M, Harun N, Liu C, Khoury JC, Broderick JP. Bayesian selective response-adaptive design using the historical control. *Stat Med*. 2018;37(26):3709-3722. doi:10.1002/sim.7836

95. 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. doi:10.1111/biom.12242

96. Ventz S, Khozin S, Louv B, et al. The design and evaluation of hybrid controlled trials that leverage external data and randomization. *Nat Commun*. 2022;13(1):5783. doi:10.1038/s41467-022-33192-1

97. Chaudhuri SE, Ho MP, Irony T, Sheldon M, Lo AW. Patient-centered clinical trials. *Drug Discov Today*. 2018;23(2):395-401. doi:10.1016/j.drudis.2017.09.016

98. 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

99. 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

100. 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

101. 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

102. Robins J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chronic Dis*. 40:139S-161S.

103. 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

104. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Med*. 2015;12(10):e1001885. doi:10.1371/journal.pmed.1001885

105. Gatto NM, Reynolds RF, Campbell UB. A structured preapproval and postapproval comparative study design framework to generate valid and transparent real-world evidence for regulatory decisions. *Clin Pharmacol Ther*. 2019;106(1):103-115. doi:10.1002/cpt.1480

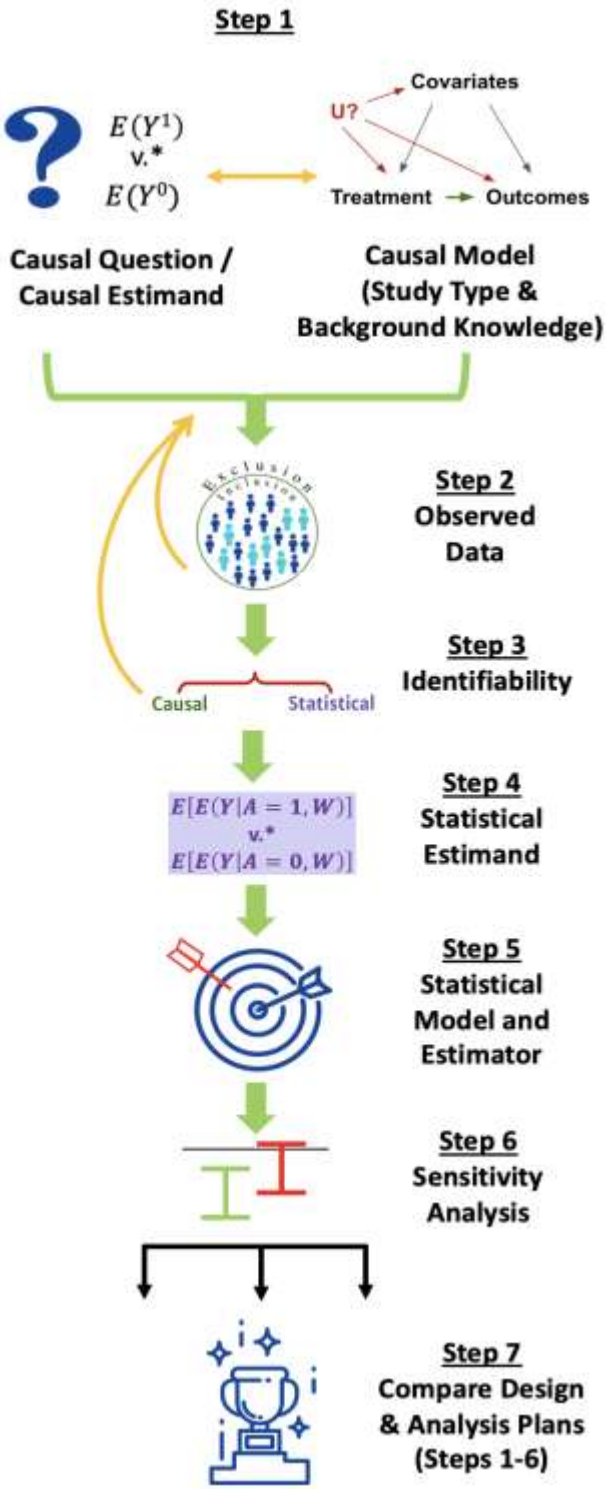
106. Bykov K, Patorno E, D'Andrea E, et al. Prevalence of avoidable and bias-inflicting methodological pitfalls in real-world studies of medication safety and effectiveness. *Clin Pharmacol Ther*. 2022;111(1):209-217. doi:10.1002/cpt.2364

107. VanderWeele TJ, Hernan MA. Causal inference under multiple versions of treatment. *J Causal Inference*. 2013;1(1):1-20. doi:10.1515/jci-2012-0002

108. Robins JM, Hernán MA. 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.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

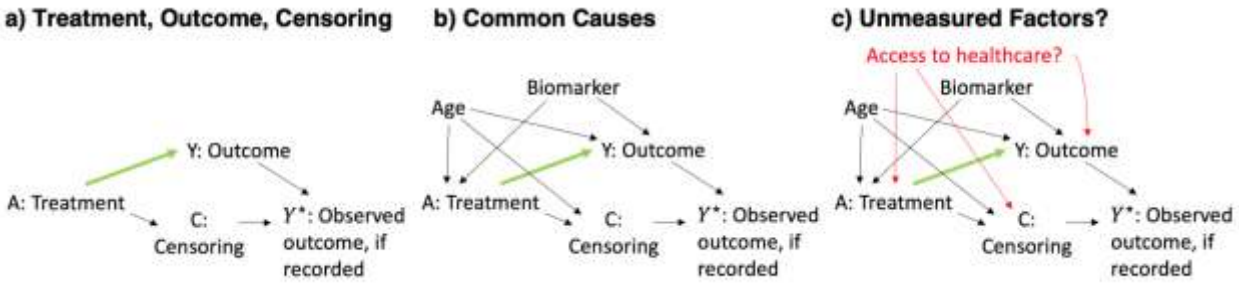
Figure 1: The *Causal Roadmap*



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

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Figure 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.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Table 1: Companion Papers Demonstrating Use of the *Roadmap*

Case Study	**Context**	**Roadmap Steps Emphasized**
Sentinel System and Scalable Phenotyping	Drug safety and monitoring	Outcome-blind[†] simulations to guide estimator pre-specification and machine learning plus natural language processing to enhance identifiability
Nifurtimox for Chagas Disease	Randomized trial infeasible	Sensitivity analysis and defining the plausible causal gap
Semaglutide and Cardiovascular Outcomes	Secondary indications	Application of the *Roadmap* to a hybrid randomized-RWD study and comparison of study designs

[†]We use outcome-blind to mean without information on the observed treatment-outcome association.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Table 2: Components of a Causal Question and Estimand per ICH E9(R1) [41] and Target Trial Emulation [17]

ICH E9(R1) attribute	**Target Trial Emulation Protocol Component**	**Explanation**	**Related Notation in this Paper**
Population	Eligibility criteria	Inclusion and exclusion criteria, including dates of eligibility, for the potential study population	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. It is also important to consider whether there are different versions of treatment (e.g., different versions of the same surgery performed by different surgeons), and which versions would be included in the treatment strategy [107].	Baseline treatment: 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 endpoint	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, causal relative risk, average treatment effect within pre-specified 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 paper we focus 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., Robins and Hernán (2009) [108], Petersen (2014) [28])

[†††]A mathematical quantity that is a function of potential outcomes (see below).

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Table 3: Examples of Identification Assumptions

Assumption	**Basic Explanation of Meaning**
Exchangeability[†]	This assumption is generally true if 1. there are no unmeasured common causes of variables that are part of the treatment strategies (Table 2: e.g., baseline or postbaseline treatment(s), censoring) and the outcome (informally, if there is no unmeasured confounding) and 2. we have not conditioned on a variable that is affected by the treatment variable(s) [32,46].
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 [32].

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

Table 4: Steps for Specifying Key Elements of a Study Design and Analysis Plan Using the *Roadmap*

Roadmap Step	

1a	• Specify the causal question and estimand.
	  • ICH E9(R1) attributes: Population, treatment, variable or endpoint, population summary [41]
	  • Target Trial Emulation Protocol Components: Eligibility criteria, treatment strategies, follow-up period, outcome, causal contrasts of interest [32]

1b	• Specify the causal model (based on background knowledge about the proposed study).
	  • Specify the type of study (e.g., traditional randomized trial, 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 will be or has been collected. • Document how the 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.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press

	● 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, confidence interval coverage of the statistical estimand, type I error probability, 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 study designs (Steps 1-6) using outcome-blind simulations based on:
	  ● causal metrics (confidence interval coverage, type I error probability, and power for a causal effect),
	  ● and metrics to quantify differences in the medical product development process of each design.
	  ● Include a comparison to a randomized trial if feasible.

https://doi.org/10.1017/cts.2023.635 Published online by Cambridge University Press