STATISTICS

A Statistical Roadmap for Journey from Real-World Data to Real-World Evidence

Yixin Fang[1] · Hongwei Wang[1] · Weili He[1]

Received: 4 June 2019 / Accepted: 17 September 2019 / Published online: 6 December 2019

Abstract

Randomized controlled clinical trials are the gold standard for evaluating the safety and efficacy of pharmaceutical drugs, but in many cases their costs, duration, limited generalizability, and ethical or technical feasibility have caused some to look for real-world studies as alternatives. On the other hand, real-world data may be much less convincing due to the lack of randomization and the presence of confounding bias. In this article, we propose a statistical roadmap to translate real-world data (RWD) to robust real-world evidence (RWE). The Food and Drug Administration (FDA) is working on guidelines, with a target to release a draft by 2021, to harmonize RWD applications and monitor the safety and effectiveness of pharmaceutical drugs using RWE. The proposed roadmap aligns with the newly released framework for FDA's RWE Program in December 2018 and we hope this statistical roadmap is useful for statisticians who are eager to embark on their journeys in the real-world research.

Keywords Causal inference · Clinical trials · Confounding bias · Statistical methods · Real-world studies

Introduction

Randomized controlled clinical trials (RCTs) are the gold standard for evaluating the safety and efficacy of pharmaceutical drugs, but their costs, duration, limited generalizability, and ethical or technical feasibility have caused some to look for real-world studies as alternatives [1]. Real-world studies include pragmatic clinical trials, single-arm clinical trials with external control, and observational studies. However, single-arm clinical trials with external control and observational studies may be much less convincing due to the lack of randomization and the presence of confounding bias [2]. Besides confounding bias, other challenges in real-world studies include addressing selection bias and measurement bias; for example, pragmatic clinical trials become observational studies due to non-compliance if the per-protocol effect is of interest [2]. The focus of this article is on addressing confounding bias, but the proposed roadmap is applicable to these other challenges.

The Food and Drug Administration (FDA) is working on ways to harmonize real-world data to create a unified system for monitoring the safety and effectiveness of medical devices and pharmaceutical drugs [3], with a target to release a draft of the guidelines by 2021. On December 7, 2018, FDA published the much-anticipated "Framework for FDA's Real-World Evidence Program" for drugs and biological products (referred to as "the framework" hereafter). The framework defines the following:

- **Real-World Data (RWD)** are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.
- **Real-World Evidence (RWE)** is the clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD.

Under the framework, evidence from RCTs will not be considered as RWE, while various hybrid or pragmatic trial designs and observational studies could generate RWE. The challenges faced by statisticians are how to generate robust RWE from the analysis of RWD. The problem is too broad to be considered in an article. Knowing that one of the main differences between RCTs and real-world studies is the presence of confounding bias, in this article, we focus on how to

✉ Yixin Fang
yixin.fang@abbvie.com

[1] AbbVie, 1 North Waukegan Rd, North Chicago, IL 60064, USA

generate robust RWE from the analysis of RWD via adjusting for confounding bias. Confounding bias is one of the three major sources of systematic bias that concern us when conducting causal inference using RWD, with the other two being selection bias and measurement bias [2]. We propose a statistical roadmap from RWD to RWE for statisticians who are eager to embark on their journey in the real-world research.

The remaining of the article is organized as follows. We first develop criteria for forming a sound research question to prepare for our journey. Next, we describe a simple setting as the departure point for our journey. For the simple setting, we propose a statistical roadmap that navigates the major steps from RWD to RWE. Then, we extend the proposed analytic plan to some more general settings. We conclude the article with a brief summary.

Methods and Results

Forming a Good Research Question

Research questions arise out of a perceived knowledge deficit within a field of study, and a good research question should be feasible, interesting, novel, ethical, and relevant (the FINER criteria [4]). Haynes developed the PICOT criteria [5], which outline five important aspects that we should consider in the development of a good research question using RCTs. The PICOT criteria cover the population (P) of interest, the intervention (I) being studied, the comparison (C) group, the outcome of interest (O), and the follow-up time (T). There are differences between RCTs and real-world studies; for example, studies that generate RWD may be non-interventional ("I" is not applicable), and assessing confounders is a critical task in observational studies ("C" could stand for confounders). To address the inapplicability of the PICOT criteria for real-world studies, it is important to develop a new set of criteria for forming a good research question using RWD.

Motivated by the PICOT criteria [5], we develop the following PROTECT criteria, which outline the five important aspects that we should consider in the development of a good research question using real-world studies. The PROTECT criteria cover the following five aspects (Table 1).

- **Population (P)** The population of patients that are of interest to the investigator; usually the population of the patients that the investigators believe will be most beneficial from the treatment of interest; the population can be determined according to a set of inclusion criteria.
- **Response/Outcome variable (R/O)** The response variable or outcome variable of a patient that the investigators intend to accomplish, measure, improve or affect; the response variable or outcome variable can be determined according to validity, reliability, and responsiveness.
- **Treatment/Exposure variable (T/E)** The treatment/exposure variable whose effect on the outcome variable is of interest to the investigators; usually the treatment/exposure variable is binary with one being the treatment of interest and zero being the treatment being compared.
- **Confounders (C)** The confounders are the variables that are associated with both the response/outcome variable and the treatment/exposure variable and controlling which will eliminate the confounding bias in estimating the causal effect of the treatment on the outcome.
- **Time (T)** The time should be considered: the time that the patient is diagnosed; the time that the patient is exposed to the treatment; the follow-up time that the outcome is assessed; and the time when the data are obtained (retrospective, cross-sectional, or prospective).

Starting with a Simple Setting

We embark on our journey by considering a simple setting where a binary outcome variable, a binary treatment variable, and a list of confounders are collected from a simple cohort study. We will describe this simple setting following the PROCTECT criteria. Then we extend the proposed strategic plan to other settings. Consider a cohort study of a specific population of patients (Population), where all the patients are exposed to either a treatment of interest or a treatment being compared (Treatment/Exposure). Assume that the treatment/exposure is time-fixed (Time), all the patients are followed by a same amount of time (Time), and the response/outcome variable is binary and measured at the end of follow-up period (Response/Outcome; Time). Assume that all the potential confounders are measured at baseline (Confounders; Time). Hereafter, we will refer this setting as "the simple setting". We emphasize that this is a simple and maybe unrealistic setting, but more complex and realistic settings will be considered later.

Table 1. The PROTECT Criteria for Forming a Research Question.

PROTECT Criteria		
P	Population/Patients	What specific pPopulation are you interested in?
R/O	Response/Outcome	What do you intend to improve or affect?
T/E	Treatment/Exposure	What is your investigational treatment/exposure?
C	Confounders	What are the potential confounders?
T	Time	What role does the time play?

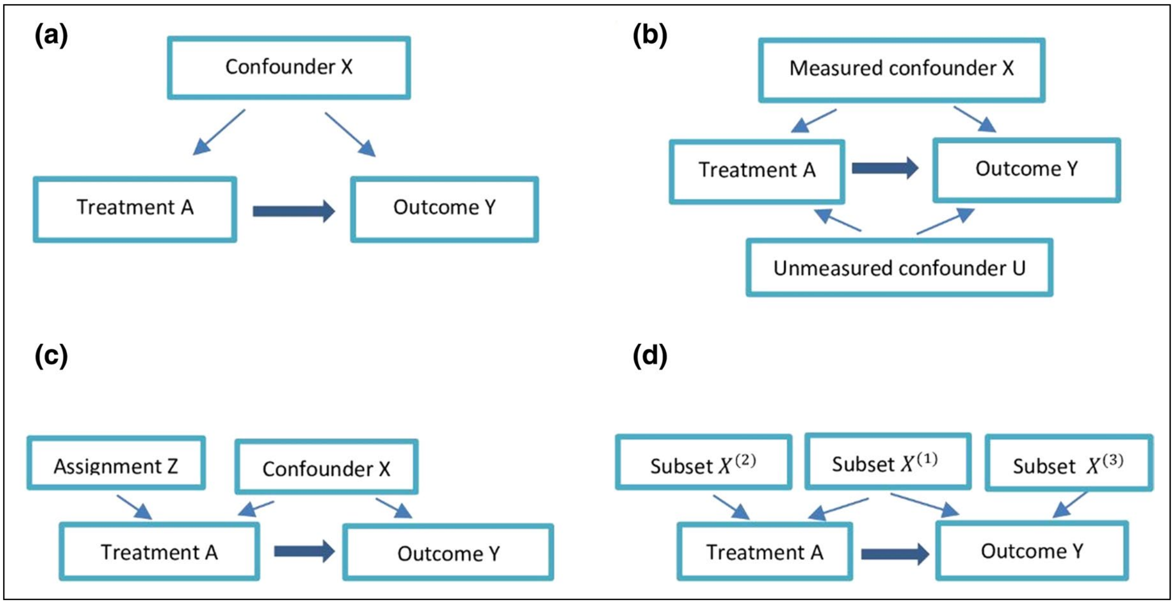

Figure 1. Directed Acyclic Graphs (DAGs) for Some Settings Under Consideration.

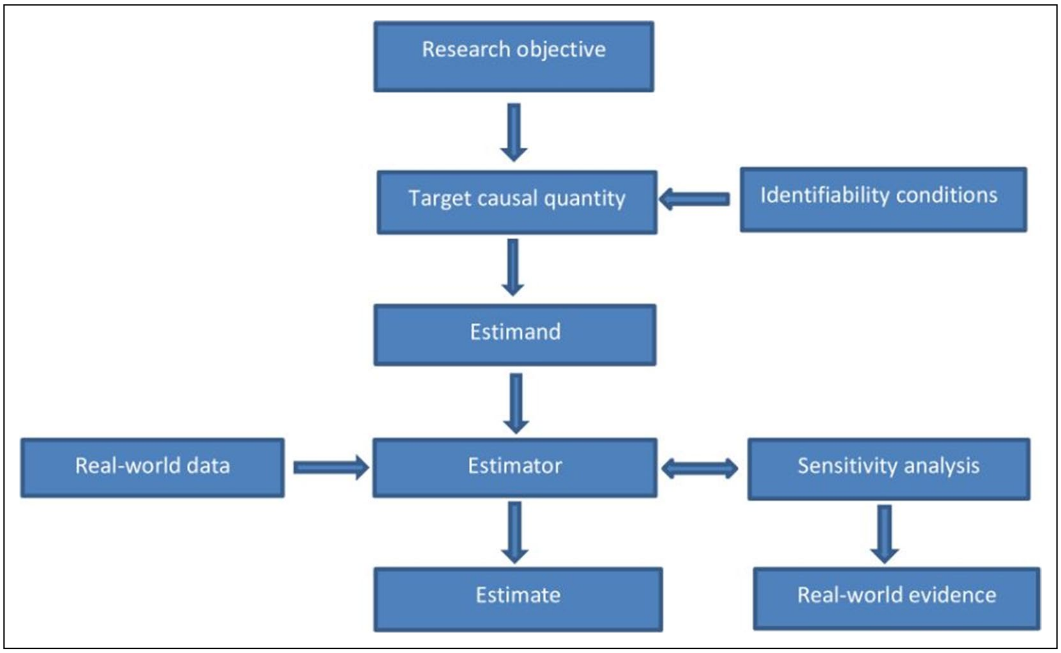

Figure 2. A Statistical Roadmap from RWD to RWE.

Let i indicate the patient ID. Denote the binary treatment variable as A_i for patient i, where $A_i = 1$ stands for that the subject is treated by the treatment of interest (for simplicity, referred to as "treated") and $A_i = 0$ stands for that the subject is treated by the treatment being compared (for simplicity, referred to as "untreated"; i.e., untreated by the treatment of interest). Denote the binary outcome variable at the end of the follow-up time T as Y_i, where $Y_i = 1$ stands for that the subject is responding to the treatment and $Y_i = 0$ stands for that the subject is not responding. Denote the vector of confounders as X_i, which may be of mixture types. Assume that there are N patients in the sample and $\left(X_i, A_i, Y_i\right),\ i = 1, \ldots, N$, are independent and identically distributed with (X, A, Y). If the response rates between the treated and the untreated are not equal, that is $P(Y = 1|A = 1) \neq P(Y = 1|A = 0)$, it is said that there is association between the outcome variable and the treatment variable. But due to the lack of randomization and the presence of confounding bias in real-world studies, as displayed in Fig. 1a, association does not imply causation. Therefore, we need to adjust for confounding bias and conduct causal inference.

A Statistical Roadmap from RWD to RWE

We propose a statistical roadmap from RWD to RWE in Fig. 2, navigating key steps in the process of generating robust RWE from the analysis of RWD. This roadmap is motivated by "*E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials*" released in 2017 by the International Council for Harmonization (ICH). Following the statistical principles that were used in ICH's E9(R1), we adapt the roadmap developed for clinical trials in E9(R1) into a roadmap that is more suitable for real-world studies. In this section, we describe key steps in the proposed roadmap for the simple setting: research objective, real-world data, targeted quantity, estimand (under identification assumptions), estimator, estimate, and sensitivity analysis (generating real-world evidence). In next section we will extend the strategic plan to other more complex settings.

Research Objective

After we form a sound research question (a perceived knowledge gap within a field of study) following the PROTECT criteria, we can develop a research objective (if achieved, would fill in the perceived knowledge gap). For the simple setting, a research objective could be to investigate the causal effect of the treatment of interest ($A = 1$) vs. the treatment being compared ($A = 0$) on the outcome variable Y measured at the end of the period T for a specific population of patients. If this objective is achieved successfully, the knowledge gap of the treatment efficacy is filled in.

Real-World Data

There are two main sources of real-world data [1]: **research data** that are collected primarily for research (e.g., Framingham Heart Study); and **transactional data** that are collected for clinical documentation (e.g., electronic health records) and administrative (e.g., claims data).

We consider those five aspects in the PROTECT criteria to assess the availability, relevance and quality of real-world data to answer the research question and the feasibility to achieve the research objective. We should select an appropriate data source that records the relevant measures ("R/O", "T/E", and "C" variables over time "T") with sufficient completeness and accuracy to study the specific population (after certain inclusion/exclusion criteria) that is specified in the research question ("P"). A typical dataset looks like the left-panel of Table 2. The ascertainment of outcome and treatment variables plays an essential role in the selection of data source. Hard clinical outcomes, such as myocardial infarction, stroke, fracture, major bleed, or death are more likely to result in valid RWD analyses [1]. For example, electronic health records capture great details of clinical outcomes and pharmacy claims capture great details of outpatient medication exposures. Moreover, the consideration of confounders and time plays an essential role in the confounding bias adjustment leading to robust RWE from RWD.

Table 2. Measured Data and Counterfactual Data of a Population of Size N.

	Measured Data			Counterfactual Data			
ID	$\boldsymbol{X_i}$	$\boldsymbol{A_i}$	$\boldsymbol{Y_i}$	$\boldsymbol{A_i = 1}$	$\boldsymbol{Y_i^1}$	$\boldsymbol{A_i = 1}$	$\boldsymbol{Y_i^0}$
1	X_1	1	Y_1	1	Y_1^1	0	Y_1^0
2	X_2	0	Y_2	1	Y_2^1	0	Y_2^0
…	…	…	…	…	…	…	…
N	X_N	1	Y_N	1	Y_N^1	0	Y_N^0

Target Causal Quantity

In order to investigate the above research objective, we should define the target causal quantity. Petersen and van der Laan made it clear that the target causal quantity (some quantity that has causal meaning and that we attempt to estimate) and the statistical estimand (some quantity that we can estimate based on the data and that is equal to the target causal quantity under certain identification assumptions) are two closely connected but different concepts [6]. In the simple setting, a good target causal quantity could be the average treatment effect of the treatment of interest vs. the treatment being compared on the outcome variable, which can be defined based on counterfactual outcomes. Recall that A is the binary treatment variable, taking on value $a = 1$ or 0. Let $Y^{a=1}$ and $Y^{a=0}$ be the outcomes that would have been observed had the patient been treated by treatment values $a = 1$ and $a = 0$, respectively. Then the average treatment effect (ATE) of the treatment on Y in the study population can be defined as the risk difference $\theta = P\left(Y^1 = 1\right) - P\left(Y^0 = 1\right)$.

In Table 2, both measured data and counterfactual data of a population of size N are displayed. The target causal quantity defined in the above (i.e., the ATE in terms of risk difference) is equal to.

$$\theta = P\left(Y^1 = 1\right) - P\left(Y^0 = 1\right)$$
$$= E\left(Y^1\right) - E\left(Y^0\right) = \frac{1}{N}\sum_{i=1}^{N} Y_i^1 - \frac{1}{N}\sum_{i=1}^{N} Y_i^0.$$

We may be interested in other types of target causal quantities. For example, the causal quantity can be defined as the relative risk in the study population, that is $P\left(Y^1 = 1\right)/P\left(Y^0 = 1\right)$.

Identifiability Conditions

The target causal quantity is defined in counterfactual outcomes. In order to translate it into some statistical estimand, which is estimable using the measured data, we need to examine the identifiability conditions. Hernan and Robins [2] stated that three identifiability conditions are needed in this simple setting for turning the target causal quantity into a statistical estimand: consistency, positivity, and conditional exchangeability.

- **Consistency** $Y^a = Y$ if $A = a$; the values of the treatment variable are well defined.
- **Positivity** $P(A = 0|X = x) > 0$ for any $X = x$ with $P(X = x) > 0$ and $a = 1$ or 0; the conditional probability of receiving each treatment is greater than zero.
- **Conditional exchangeability** $\mathrm{Y}^{\mathrm{a}} \perp\!\!\!\perp A|X$ for $a = 1$ or 0; the conditional probability of receiving each treatment depends only on the measured confounders. This is equivalent to *the assumption of no unmeasured confounding*.

The assumption of no unmeasured confounding is displayed in a DAG in Fig. 1a.

Estimand

Under the above three identifiability conditions, we can translate the target causal quantity into a statistical estimand. Consider the target causal quantity $\theta = E\left(Y^1\right) - E\left(Y^0\right)$, which is not estimable because it depends on counterfactual outcomes, Y^1 and Y^0, only one of which is measured for each patient. Under the identifiability conditions, we have the following equation:

$$\theta = E\left(Y^1\right) - E\left(Y^0\right) = E\left\{E\left(Y^1|X\right) - E\left(Y^0|X\right)\right\}$$
$$= E\{E(Y^1|A = 1, X) - E\left(Y^0|A = 0, X\right)\}$$
$$= E\{E(Y|A = 1, X) - E(Y|A = 0, X)\},$$

where the first equality is the definition of the causal quantity under consideration, the second uses the law of total expectation, the third equality uses the conditional exchangeability condition, and the fourth equality uses the consistency condition. Then the rightest term of the above equation

$$\theta^* = E\{E(Y|A = 1, X) - E(Y|A = 0, X)\}$$

is referred to as the statistical estimand, which is estimable and is equal to the target causal quantity $\theta = E\left(Y^1\right) - E\left(Y^0\right)$ under the identification conditions. Note that θ^* is estimable because it depends on measured variables Y, A and X. To summarize, this statistical estimand has following three properties:

- It is a parameter associated with the population;
- It is equal to the target causal quantity under certain identifiable conditions;
- It is estimable using the measured data under certain identifiable conditions.

Estimator

There are a variety of statistical estimation methods which can be applied to estimate the statistical estimand $\theta^* = E\{E(Y|A = 1, X) - E(Y|A = 0, X)\}$. The resulting quantity from each estimation method for estimating the estimand is referred to as an estimator. Based on our literature review, we can loosely categorize these methods into three categories: (1) Stratification-based methods [7]; (2) G-methods [2]; and (3) Targeted learning methods [8, 9].

- **Stratification-based methods** This category includes stratification, restriction, and matching. Usually stratification and matching are based on propensity scores [10]. These methods work well in practice for the simple setting because they are robust and easy to implement and interpret. These methods usually require specifying a statistical model for the treatment variable against confounders; that is a model of $A \sim X$. In addition, these methods may not be generalized to complex longitudinal cohort studies in the presence of time-varying treatments and time-varying confounding.
- **G-methods** This category includes g-formula, inverse-probability (IP) weighting, and augmented IP weighting [2]. These methods are able to be generalized to longitudinal cohort studies in the presence of time-varying confounding and this is reason they are referred to as g-methods, where "g" stands for "generalized". These methods also require specifying some statistical model: a model for the outcome $Y \sim A + X$ in g-formula, a model for the treatment $A \sim X$ in IP weighting, and models for both $Y \sim A + X$ and $A \sim X$ in augmented IP weighting. The g-formula estimator is asymptotically unbiased only if the outcome model is estimated consistently; the IP

weighting estimator is asymptotically unbiased only if the treatment model is estimated consistently; and the augmented IP weighting estimator is asymptotically unbiased if either the outcome model or the treatment model is estimated consistently (i.e., doubly robust).

- **Targeted learning methods** The core of this category of targeted learning methods is targeted maximum likelihood estimators (TMLEs), where "targeted" means that the methods attempt to estimate the target causal quantity θ and "MLE" indicates that the methods are efficient (roughly speaking, MLE provides an asymptotically minimum-variance unbiased estimator for the statistical estimand θ^*). TMLEs are based on the g-formula and therefore they are also g-method that can be generalized to complex longitudinal cohort studies in the presence of time-varying confounding. Compared with the classical g-methods in the previous category, TMLEs have more desirable asymptotic properties such as consistency without specifying parametric models (super learning plays an important role in the construction of TMLEs [11]) and efficiency.

Estimate

An estimator is a function of data. When the data are observed and are plugged into the function, it produces a point estimate. Along with some measurement of the uncertainty (e.g., standard error), we can conduct statistical inference.

We rely on powerful computing tools such as SAS procedures and R packages to calculate the estimates from the estimators, using the observed real-world data.

- To implement the stratification-based methods, for example, stratification and matching based on propensity scores, we can use **SAS procedure PSMATCH**;
- To implement the g-methods, including g-formula, IP weighting, and augmented IP weighting, we can use **SAS procedure CAUSALTRT**;
- To implement the targeted learning methods, we can use **R package tmle** for the simple setting and **R package ltmle** for longitudinal cohort studies.

Sensitivity Analysis

Recall that in the translation from the target causal quantity to the statistical estimand, which in turn is to be estimated by an estimator, we make three identifiability conditions. However, these conditions, particularly the conditional exchangeability condition (which is equivalent to the assumption of no unmeasured confounding) cannot be tested using the measured data. Therefore, we should conduct sensitivity analysis to evaluate how the estimate would change if the assumption of no unmeasured confounding were violated. Figure 1b shows the DAG for the above simple setting with unmeasured confounder U.

We can use Monte Carlo simulation to conduct sensitivity analysis as to that the assumption of no unmeasured confounding is violated. To conduct Monte Carlo sensitivity analysis, we can use statistical software to generate some bias parameters as the unmeasured confounder U and then invert these bias parameters to provide a distribution of bias-corrected estimates. Since the first formal sensitivity analysis to detect unmeasured confounding bias was published [12], there have been many applications of sensitivity analysis, as comprehensively reviewed by Greenland [13].

Real-World Evidence

The most recent development of sensitivity analysis is E-value proposed by VanderWeele and Ding [14], where "E" stands for "evidence". As pointed out by VanderWeele and Ding [14], *P value gives evidence for association and E-value gives evidence that the association is causation.* Simply put, E-value measures the real-world evidence that is derived from the analysis of real-world data. This makes the destination of our journey from RWD to RWE.

To introduce E-value, denote the estimated relative risk of A on Y as RR_{AY}, after adjusting for measured confounding via some estimation method, say targeted learning. In addition, denote the relative risk of U on A as RR_{UA} and the relative risk of U on Y as RR_{UY}. Under the assumption of no unmeasured confounding, RR_{AY} is an asymptotically unbiased estimator of the average treatment effect of A on Y (here we consider the target causal quantity in terms of relative risk $\theta = P\left(Y^1 = 1\right)/P\left(Y^0 = 1\right)$). Assume that we wish $\mathrm{RR}_{AY} > 1$, which means that the treatment of interest $A = 1$ increases the response rate. However, in the presence of unmeasured confounding U, RR_{AY} is a biased estimate the target causal quantity θ. VanderWeele and Ding [14] showed that the bias due to the presence of unmeasured confounding equals

$$B = \frac{\mathrm{RR}_{UY} \times \mathrm{RR}_{UA}}{\mathrm{RR}_{UY} + \mathrm{RR}_{UA} - 1}.$$

If RR_{UA} and RR_{UY} were known, we could shift the estimator to $\frac{\mathrm{RR}_{AY}}{B}$, along with shifting its confidence interval, to adjust for the unmeasured confounding U. But since RR_{UA} and RR_{UY} are unknown, we can evaluate what values that RR_{UA} and RR_{UY} take on will lead to the disappearance of causation founding. VanderWeele and Ding [14] showed that if RR_{UA} and RR_{UY} were to be greater than

$$\text{E-value} = \text{RR}_{AY} + \left[\text{RR}_{AY} \times \left(\text{RR}_{AY} - 1\right)\right]^{1/2},$$

the causation founding would disappear; that is, $\text{RR}_{AY} > 1$ would become $\text{RR}_{AY}/B \leq 1$.

VanderWeele and Ding [14] illustrated E-value using the famous example of the association between smoking and lung cancer. Hammond and Horn [15] obtained the estimated relative risk after adjusting for measured confounders $\text{RR}_{AY} = 10.73$ (95% CI 8.02, 14.36). But Fisher [16] thought the smoking-lung cancer relationship could be explained by a genetic variant U. VanderWeele and Ding [14] provided that E-value for the estimate is 20.9 and E-value for the lower bound of the CI is 15.5, showing that the evidence that the association is causation is very strong. With an observed relative risk of $\text{RR}_{AY} = 10.73$, an unmeasured confounder that was associated with both the outcome and the exposure by a relative risk of 20.9-fold each, above and beyond the measured confounders, could explain away the estimate, but weaker confounding could not. We can use **R package EValue** to calculate E-value.

Discussion: Extension to Other Settings

We have proposed a statistical roadmap from RWD to RWE, focusing on the simple setting. We will extend the proposed strategic plan to other settings. We describe these other settings categorized according to the five aspects of the PROTECT criteria.

Population

In the simple setting, we define the target causal quantity as the average causal effect of the treatment variable on the outcome variable over the entire study population. In other settings, the target causal quantity may be defined as the average causal effect of the treatment variable on the outcome variable in some subset of the population, say, in the patients that are treated by the treatment of interest, i.e., $E\left(Y^1|A = 1\right) - E\left(Y^0|A = 1\right)$, which is referred to as the average treatment effect on the treated (ATT).

For example, in non-randomized, single-arm clinical trials with external RWD control, where $A = 1$ denote the treated arm and $A = 0$ denote the control arm. We are interested in the average causal effect of the treatment on the outcome in the treated patients only. Assuming those three identifiability conditions, the statistical estimand for ATT is $E_{X|A=1}[E(Y|A = 1, X) - E(Y|A = 0, X)]$, where the outer expectation is over the conditional distribution of X given $A = 1$.

Response/Outcome Variable

In the simple setting, we consider binary outcome variable. In other settings, the outcome variable can be continuous or time-to-event. If the outcome variable is continuous, the strategic plan is similar to the one for the simple setting because we can consider the target causal quantity in terms of mean difference, $E\left(Y^1 - Y^0\right)$, with estimand $E[E(Y|A = 1, X] - E(Y|A = 0, X)]$.

If the outcome variable is time-to-event (a.k.a. survival outcome), due to the built-in selection bias of hazard ratios [2], we usually do not define the target causal quantity in terms of hazard ratio; instead, we define the target causal quantity in terms of survival rate. Let $Y^{a,0}$ denote the survival time if the patient were treated by the treatment $A = a$ and were not censored. Then we can consider the difference in the survival rates as the target causal quantity; i.e., $P\left(Y^{1,0} > t\right) - P(Y^{0,0} > t)$, for any given $t > 0$.

Treatment/Exposure Variable

In the simple setting, we assume there is no non-compliance (and, in general, no measurement error); that is, the consistency condition is satisfied. In other settings, we may consider the impact of non-compliance. Non-compliance is one type of measurement errors, leading to measurement bias [2].

For example, in pragmatic randomized clinical trials, although the treatments are assigned to patients randomly, the patients may not follow the assignments in real world. Figure 1c displays a DAG for such setting, where Z is the random assignment that a patient receives and A is the actual treatment that the patient takes. If we are interested in the intention-to-treat effect, then we can define the target causal quantity as $E\left(Y^{Z=1} - Y^{Z=0}\right)$ and there is no confounding thanks to the randomization in Z. If we are interested in the per-protocol effect, then we can define the target causal quantity as $E\left(Y^{A=1} - Y^{A=0}\right)$ and we need to worry about confounding. Therefore, if we are interested in the per-protocol effect, we should view the randomized clinical trial as an observational study [2].

Confounders

In the simple setting, the vector of potential confounders is not high-dimensional; that is, the positivity condition is satisfied. In other settings, the vector of potential confounders may be high-dimensional and the positivity condition may be violated. Therefore, we should take variable selection into account in the adjustment of confounding.

For the settings where there are too many confounders, as in Fig. 1d, we can divide the set of possible confounders into three subsets, $X = X^{(1)} \cup X^{(2)} \cup X^{(3)}$, where variables in $X^{(1)}$ are truly confounders, variables in $X^{(2)}$ are only associated with treatment but not outcome, and variables in $X^{(3)}$ are only associated with outcome but not treatment. By extensive simulations [17], Brookhart et al. suggested that we should include variables in $X^{(1)} \cup X^{(3)}$ and exclude variables in $X^{(2)}$ in the adjustment of confounding. A bad practice is to use some variable selection procedures to select variables that predict the treatment variable, leading to include variables in $X^{(1)} \cup X^{(2)}$ but exclude $X^{(3)}$. We can apply the *collaborative targeted learning* (C-TMLE) for such settings in the adjustment of confounding [8, 9].

Time

Time plays an important role in the selection of RWD. In the simple setting, we assume that all the patients are followed from the baseline by the same period of time T, all the confounders are measured at baseline, the outcome variable is measured at the end of T for each patient, and the treatment variable stays the same from baseline during the follow-up period. In pragmatic randomized clinical trials and single-arm clinical trials with external RWD control, we wish these ideal assumptions are satisfied. If not, we should define the target causal quantity in terms of time-varying treatment.

For observational studies, we should consider the role of time in the design stage. Figure 3 displays the role of time in some commonly used observational study designs. In case–control studies, the diseased cases and the disease-free controls are matched based on some measured confounders, and the exposure variables in the past are recalled by the patients. Case–control studies suffer from measurement bias (due to the recall bias) and confounding bias (due to confounders that are not used for matching). To adjust for confounding in case–control studies or cross-sectional studies, the strategic plan is similar to the one used for the simple setting.

However, for longitudinal cohort studies (including retrospective cohort study and prospective cohort study) with time-varying treatment, the situation becomes very complex. We refer them as complex longitudinal studies [9]. We will follow the same roadmap from RWD to RWE, but subtle details should be added to the strategic plan in the aspects of causal quantity, identifiability conditions, and estimand. The remaining of this subsection mainly comes from Part III of Hernan and Robins [2].

Consider a time-varying binary treatment A_k that may change at every follow-up visit indexed by k, where $k = 0, 1, \ldots, K$, with 0 being the baseline and K the last follow-up visit. Let $\bar{A}_k = \left(A_0, A_1, \ldots, A_k\right)$ denote the treatment history from baseline to follow-up visit k, and let $\bar{A} = \bar{A}_K$ denote the entire treatment history through K follow-up visits. For example, two treatment strategies are "always treat" and "never treat", represented by $\bar{a} = (1, \ldots, 1) = \bar{1}$

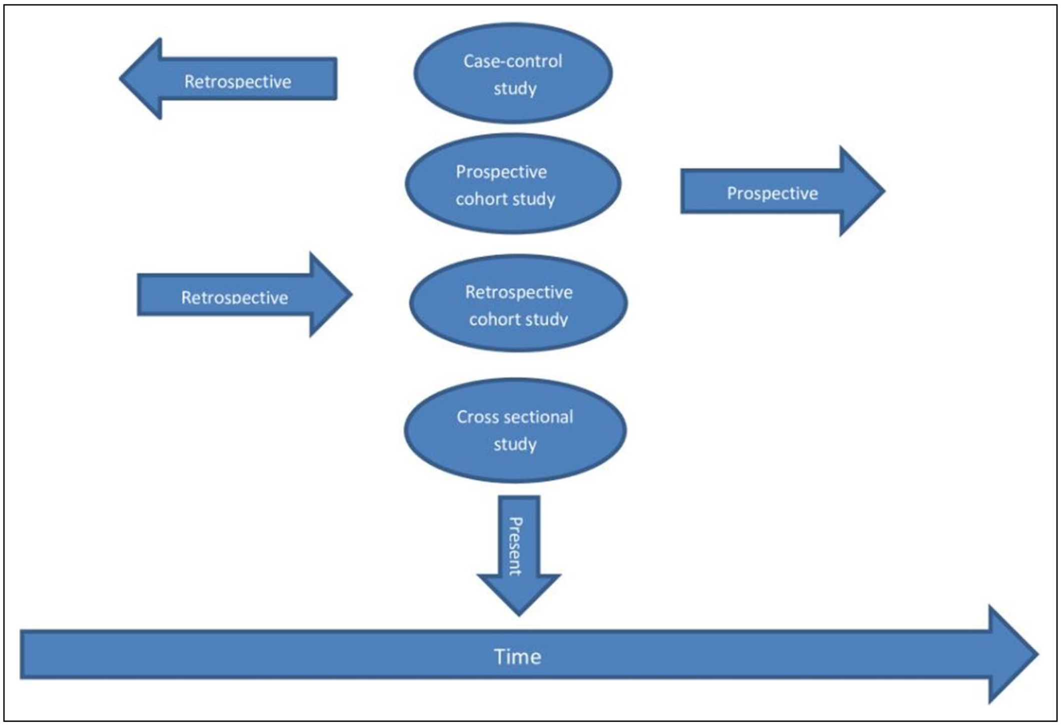

Figure 3. The Role of Time in Observational Study Designs.

and $\bar{a} = (0, \ldots, 0) = \bar{0}$. We can define the target causal quantity as the average causal effect of $\bar{A}$ on the outcome Y; that is, the contrast between the mean counterfactual outcome $Y^{\bar{a}=\bar{1}}$ and the mean counterfactual outcome $Y^{\bar{a}=\bar{0}}$, i.e., $E\left[Y^{\bar{a}=\bar{1}} - Y^{\bar{a}=\bar{0}}\right]$.

Consider a vector of time-varying covariates X_k measured at every follow-up visit k, where $k = 0, 1, \ldots, K$. Assume the variables are observed in the following sequential order, $X_0 \to A_0 \to X_1 \to A_1 \to X_K \to A_K \to Y$. Let $\bar{X}_k = \left(X_0, X_1, \ldots, X_k\right)$ denote the covariates history from baseline to follow-up visit k. There are two types of treatment strategies, static treatment strategies in which treatment does not depend on covariates and dynamic treatment strategies in which the treatment $A_k = a_k$ depends on the evolution of the patient's time-varying covariates $\bar{X}_k$. In order to develop the statistical estimand associated with the target causal quantity, we need to make some identifiability conditions (sequential consistency, sequential positivity, and sequential exchangeability), which can be examined by sensitivity analysis discussed in Sects. 3.8–3.9.

Then we can develop the statistical estimand associated with the target causal quantity. For example, if the target causal quantity is $E\left[Y^{\bar{a}=\bar{1}} - Y^{\bar{a}=\bar{0}}\right]$, under those identifiability conditions, the estimand is $E[E\left(Y|\bar{A} = \bar{1}, \bar{X}\right) - E(Y|\bar{A} = \bar{0}, \bar{X})]$. Using RWD, the estimand can be estimated by either the g-methods [2] or the targeted learning methods [9].

Conclusion

In this article, we suggest the PROTECT criteria for developing a sound research question in real-world studies. Then we propose a roadmap from real-world data to real-world evidence, in alignment with the recently released framework for FDA's Real-World Evidence Program. The proposal is based on extensive review of the literature of causal inference that is relevant to the adjustment for confounding bias. We first describe the roadmap under a simple setting and then extend the strategic plan to other settings. Those other settings are categorized according to the five aspects of the PROTECT criteria.

Disclaimers

The comments provided here are solely those of the presenters and are not necessarily reflective of the positions, policies or practices of authors' employers.

Compliance with Ethical Standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? *Clin Pharmacol Ther*. 2017;102(6):924–33.
2. Hernan MA, Robins JM. *Causal inference: What If*. Boca Raton: Chapman & Hall/CRC; 2020.
3. Sherman RE, Anderson SA, Pan GJD, et al. Real-world evidence—what is it and what can it tell us? *N Engl J Med*. 2016;375(23):2293–7.
4. Farrugia P, Petrisor BA, Farrokhyar F, Bhandari M. Research questions, hypotheses and objectives. *Can J Surg*. 2010;53(4):278–81.
5. Haynes RB. Forming research questions. *J Clin Epidemiol*. 2006;59(9):881–6.
6. Petersen ML, van der Laan MJ. Causal models and learning from data. *Epidemiol Camb Mass*. 2014;25(3):418–26.
7. Imbens GW, Rubin DB. *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press; 2015.
8. van der Laan MJ, Rose S. *Targeted learning: causal inference for observational and experimental data*. New York: Springer; 2011. http://public.eblib.com/choice/publicfullrecord.aspx?p=763456. Accessed May 17, 2019.
9. van der Laan MJ, Rose S. *Targeted learning in data science: causal inference for complex longitudinal studies*. New York: Springer; 2018.
10. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
11. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007. https://doi.org/10.2202/1544-6115.1309.
12. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. *J Natl Cancer Inst*. 1959;22(1):173–203.
13. Greenland S. Multiple-bias modelling for analysis of observational data. *J R Stat Soc Ser A*. 2005;168(2):267–306.
14. VanderWeele T, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med*. 2017;167(4):268–74.
15. Hammond EC, Horn D. Smoking and death rates—report on 44 months of follow-up of 187,783 men: 2. Death rates by cause. *Am Med Assoc*. 1958;166(11):1294–308.
16. Fisher RA. Cancer and smoking. *Nature*. 1958;182:596.
17. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149–56.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.