Sensitivity analysis for causality in observational studies for regulatory science

Iván Díaz[1] ●, Hana Lee[2], Emre Kıcıman[3], Edward J. Schenck[4], Mouna Akacha[5], Dean Follman[6] and Debashis Ghosh[7]

[1]Division of Biostatistics, Department of Population Health, New York University Grossman School of Medicine, New York, NY, USA; [2]Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA; [3]Microsoft Research, Redmond, WA, USA; [4]Department of Medicine, Weill Cornell Medicine, New York, NY, USA; [5]Novartis Pharma AG, Basel, Switzerland; [6]Biostatistics Research Branch, National Institute of Allergy and Infectious Disease, Silver Spring, MD, USA and [7]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Colorado, USA

Abstract

Objective: The United States Congress passed the 21st Century Cures Act mandating the development of Food and Drug Administration guidance on regulatory use of real-world evidence. The Forum on the Integration of Observational and Randomized Data conducted a meeting with various stakeholder groups to build consensus around best practices for the use of real-world data (RWD) to support regulatory science. Our companion paper describes in detail the context and discussion of the meeting, which includes a recommendation to use a causal roadmap for study designs using RWD. This article discusses one step of the roadmap: the specification of a sensitivity analysis for testing robustness to violations of causal model assumptions. *Methods:* We present an example of a sensitivity analysis from a RWD study on the effectiveness of Nifurtimox in treating Chagas disease, and an overview of various methods, emphasizing practical considerations on their use for regulatory purposes. *Results:* Sensitivity analyses must be accompanied by careful design of other aspects of the causal roadmap. Their prespecification is crucial to avoid wrong conclusions due to researcher degrees of freedom. Sensitivity analysis methods require auxiliary information to produce meaningful conclusions; it is important that they have at least two properties: the validity of the conclusions does not rely on unverifiable assumptions, and the auxiliary information required by the method is learnable from the corpus of current scientific knowledge. *Conclusions:* Prespecified and assumption-lean sensitivity analyses are a crucial tool that can strengthen the validity and trustworthiness of effectiveness conclusions for regulatory science.

Introduction

Real-world data (RWD), such as administrative claim records, electronic health records, and large registries, provide unprecedented quantities of data on millions of patients and thousands of variables in real-world settings. As such, RWD constitute an extraordinary opportunity to generate practice-based evidence to improve healthcare and health outcomes, so-called real-world evidence (RWE). Recognizing the importance of RWE for regulatory purposes, the United States Congress passed the 21st Century Cures Act [1] that mandated the development of United States Food and Drug Administration (FDA) guidance on regulatory use of RWE to support regulatory decisions. Despite the many potential advantages, the prospect of incorrect effect estimates has historically cast doubt on the use of RWE for regulatory science. Indeed, the principle that "correlation does not imply causation" is a fundamental concept used across various scientific fields to prevent logical fallacies and erroneous scientific conclusions, which are rightfully central to most criticisms of using RWD for regulatory science.

However, scientists frequently gain knowledge about cause and effect based on statistical associations. For instance, a statistical association may be interpreted as a causal relationship when it is known that there is no unmeasured confounding, and the direction (e.g., time-ordering) of the causal relationship is already known. One can make such strong assumptions given external knowledge, for example, that data come from a perfectly executed randomized study with no loss-to-follow-up and perfect adherence. Broadly speaking, causal interpretation must be supported by external knowledge of the data-generating process, such as study design or mechanistic knowledge about the phenomena under investigation. This external knowledge is often encoded in a *causal model,* and the set of models and data analysis tools concerned with the appropriateness of such causal interpretations is known as *causal inference* (please see our companion paper [2] on *the causal roadmap* for a more detailed discussion on causal models and causal inference).

Positing causal models with RWD involves making non-testable assumptions, such as assuming the absence of unmeasured confounding variables, time-ordering between the variables, no adjustment for colliders, monotonicity for instrumental variables, etc. Absence of unmeasured confounding is an important assumption that must primarily be addressed at the causal model stage by making every effort to posit a causal model that corresponds to the state-of-the-art in the substantive field and by making every effort to measure all confounders dictated by the model. For instance, RWD analyses seeking to establish the effectiveness of COVID-19 vaccines for the prevention of Post-Acute Sequelae of COVID (PASC) require understanding and measuring all the patient characteristics that lead patients to get vaccinated in the real world, as well as whether they are likely to affect the risk of developing PASC. However, despite best efforts, there may be situations where the causal model is incorrect, or where some confounders are unmeasurable with current technology or available data. For instance, in an analysis based on Electronic Health Records, certain important socioeconomic factors that may confound the vaccination-PASC relation may be unmeasured. In such cases, the statistical parameter targeted by the analysis may not have a causal interpretation. The use of RWD for regulatory science requires maximum efforts to ensure dependable causal inferences, even when the assumptions of the causal model are incorrect. In the context of plausible violations to the assumptions of the causal model, or the inability to measure some of the confounders dictated by the model, sensitivity analyses are a valuable tool that can be used to make more dependable causal inferences from RWD.

While we often cannot validate an untestable assumption, we can often test how sensitive our scientific conclusions are to violations of our assumptions. To this end, we use a *sensitivity parameter* which encodes the severity of violations to the assumptions of the model, with the goal of determining if the maximum sensible value of the sensitivity parameter (which should be prespecified, as discussed below) is large enough to invalidate the scientific conclusions derived from adjusted statistical estimates. This simple but powerful idea has a long-standing history in epidemiological sciences and is currently part of the International Council for Harmonization E9 Guidance on Statistical Principles for Clinical Trial [3,4]. One of the most well-known examples is its application in 1959 by Cornfield *et al.* [5] who demonstrated that if an unmeasured confounder can explain the observed association between smoking and lung cancer, it would need to cause a nine-fold increase in the probability of smoking. Multiple attempts were made to find such a strong confounder, but all such conjectured confounders (e.g., genetic, hormonal) had an effect on smoking that was much lower than the nine-fold increase necessary to invalidate causal conclusions. As a result, Cornfield et al. concluded that smoking causes lung cancer. This analysis played a pivotal role in establishing a public consensus about the causal relationship between smoking and lung cancer [6]. Others arrived at qualitatively similar conclusions using alternative sensitivity analyses [7].

The smoking and lung cancer example is a "success" story in the sense that it exemplifies a case where sensitivity analyses prove that an observed association is causal. Perhaps more importantly, sensitivity analyses can be used in the opposite direction to unveil cases where unmeasured confounding could easily explain away an observed association. An example is the effect of hormone replacement therapy (HRT) on cardiovascular disease (CVD), where multiple observational studies showed that HRT reduced the risk of CVD [8,9], but subsequent randomized trials demonstrated that in fact HRT increases the risk of CVD [10]. If the original observational studies had conducted a sensitivity analysis, they would have found that an unmeasured confounder with a weak association with the exposure (odds ratio 1.13) would have been sufficient to explain away the observed protective association [11], although it is worth noting that some controversy remains about the effect of HRT [12].

Before we proceed, it is important to clarify that we refer to sensitivity analyses as methodologies that aid in testing the extent to which varying violations of *causal modeling assumptions* would lead to different conclusions. This kind of sensitivity analysis must be distinguished from analyses that seek to test the extent to which *statistical modeling assumptions* would lead to different conclusions. Statistical and causal sensitivity analyses are fundamentally different in that the former seeks to assess the validity of testable assumptions, whereas the latter seeks to assess the validity of untestable assumptions. For instance, goodness of fit of a logistic regression model may be tested by assessing predictive accuracy after adding additional terms or comparing to other regression models. In contrast, it is impossible to learn from data whether we have measured all the relevant confounders, or whether some of the variables that we are adjusting for are not confounders but are colliders and therefore induce bias. Causal modeling assumptions must be therefore supported based on background substantive information and, when doubted, must be tested with an appropriate sensitivity analysis.

The objective of sensitivity analysis may be simple, but the methods used to express violations of model assumptions, to define sensitivity parameters, and to test their magnitude can be complex. In this article, we provide a brief review of various methods for sensitivity analysis and demonstrate their usefulness in using RWD to establish causality to support regulatory submissions. We begin with a case study that presents an observational analysis of the effectiveness of Nifurtimox (NFX), a medication for the treatment of Chagas disease. We then proceed with a review of the most common methods for sensitivity analysis and conclude with recommendations for their use in supporting regulatory submissions.

Case study: The effectiveness of Nifurtimox in the treatment of the Chagas disease

Background on the Chagas disease

American *trypanosomiasis*, also called Chagas disease, is caused by the parasite *Trypanosoma cruzi.*, which is transmitted by an insect vector. The disease affects around 8 to 10 million people in the endemic zones of Latin America, from the South of the United States of America(USA) to the North of Argentina. Although the disease was traditionally restricted to Latin America, a growing number of cases have been reported in the USA. Today, the disease is classified as one of the leading neglected tropical diseases in the USA [13], with up to 350,000 persons infected. *T. cruzi* is transmitted by the bite of several species of hematophagous bugs. The parasites are excreted in the feces of the bugs and penetrate human hosts through the mucosa or through scratches in the skin. After localized multiplication, the parasite is then dispersed to target organs (principally the intestinal or cardiac nerve plexus) through invasion of the bloodstream. The acute phase following infection lasts 4-6 weeks and is generally asymptomatic but may lead to fever, malaise, myalgia, and headaches. In more than

Table 1. Number of patients in the treated and control group according to their outcome and censoring status.

	NFX	Control
Seroreversion	16	1
No seroreversion	3	27
Lost to follow-up	36	321
Total	55	349

one-third of chronically infected individuals, clinical disease reappears after a period of latency lasting between 10 and 30 years. The chronic stage of the disease manifests as irreversible lesions mainly affecting the cardiac and digestive systems. The chronic form is also associated with a risk of sudden death. Diagnosis is made following detection of trypanosomes in the blood in the acute phase or through serological testing which detects antibodies made to fight the trypanosome infection [14].

Nifurtimox is one of the drugs currently used in endemic areas of Latin America to treat the Chagas disease. Despite the public importance of the disease, Nifurtimox is currently not approved by the FDA for adults, partly because few research studies exist about its efficacy. Nifurtimox was first approved in the USA for the treatment of Chagas disease in pediatric patients on the basis of the results of a randomized study that established the effect of the drug to induce negative seroreversion or seroreduction $\geq 20\%$ one year after treatment [15]. The long incubation periods of the disease (up to 30 years) mean that the cost of a randomized study to assess the effectiveness of Nifurtimox in the full long-term span of the disease is prohibitive.

Data source

Few studies, randomized or otherwise, exist that follow groups of patients over such long periods of time and provide a proper long-term account of the clinical efficacy of treatment with Nifurtimox. One such study, conducted by Fabbro *et al.* [16], followed a group of 404 patients recruited between 1976 and 1999. Data from this study had the following problems which made them not immediately usable for assessing the effectiveness of Nifurtimox:

1. Treatment was assigned mostly based on availability, patients' willingness to be treated, often considering the baseline health status of the patient. Consequently, a naive analysis of the data that does not adjust for these confounders will result in biased inference.

2. Since some patients were lost to follow-up during the study, outcome data are subject to informative missingness. If the reasons why patients were lost to follow-up during the study are related to the outcome of interest (e.g., patients lost to follow-up were because of their health status, ignoring that information will also result in biased inference.

The outcome of interest in this study is negative seroconversion 30 years after treatment (henceforth referred to as seroreversion), meaning that no evidence of presence of the parasite remains in serological blood tests. Due to the long study period, there is substantial loss-to-follow-up. Table 1 presents the distribution of the outcome across treatment groups in the study.

The potential for bias is clear from this table. With over 90% of observations lost to follow-up in the control group, it may initially seem impossible to use this data to assess the effectiveness of NFX on 30-year seroreversion without bias.

To overcome the initial barrier of large loss-to-follow-up rates, it is possible to consider external information, such as the small rate of seroreversion when patients are untreated (henceforth referred to as *spontaneous seroreversion*). For instance, two studies in children report a rate of about 5% [17,18], while a meta-analysis of studies in adults reports a rate as low as 2% [19]. The significance of these low rates becomes clear when compared to the most conservative imputation strategy for the missing NFX patients. If all 36 lost to follow-up NFX patients did not serorevert, the resulting NFX seroreversion rate would be 16 out of 55, or 29%. This is considerably higher than the externally supported rate of 5% for spontaneous seroreversion. However, even if the rate of spontaneous seroreversion is as high as 10, 15%, or 20%, the data still support the hypothesis that NFX induces seroreversion in Chagas patients.

In what follows we use this dataset as an illustrative example for how to conduct a sensitivity analysis, keeping in mind that regulatory decision-making also relies on multiple additional issues such as whether the data are "fit-for-purpose [20]," which we do not address here. The analysis is based on using the rate of spontaneous seroreversion as a *sensitivity parameter*, where the conclusions of effectiveness of NFX are assessed in light of various plausible values of this sensitivity parameter.

Nonparametric methodology for sensitivity analysis using rates of seroreversion as a sensitivity parameter

The above ideas may be formalized in a rigorous statistical procedure for sensitivity analysis as follows. First, consider a target estimand of interest defined as the *average treatment effect on the treated*, $\psi^c = E[Y(1)-Y(0)|A=1]$, where $A = 1$ denotes treatment with NFX and $A = 0$ denotes control, $Y(1)$ denotes the potential 30-year seroreversion status of a patient if, possibly contrary to fact, they were treated with NFX, $Y(0)$ denotes the potential 30-year seroreversion status of a patient if, possibly contrary to fact, they were untreated, and $E[Y(1)-Y(0)|A=1]$ denotes taking the expectation (mean) of the difference between potential outcomes in the population of treated patients. The parameter ψ^c is the *target causal estimand*, interpreted as the difference in outcome rates among treated patients in hypothetical worlds where NFX was given to all vs no Chagas patients. If we knew this number, we would know whether NFX induces higher rates of seroreversion in the patients who are treated with it. Without further assumptions, this quantity is not estimable since we cannot possibly observe a patient's outcome under treatment *and* under no treatment.

In addition to the data on treatment, seroreversion, and loss-to-follow-up, Fabbro et al. collected multiple important baseline variables on the patients in the cohort, including age, sex, initial serology titers, as well as the presence of Chagas-related abnormalities in the electrocardiogram. We use the letter W to denote a vector containing these variables and use $C = 1$ to indicate that a patient had complete follow-up and the study endpoint was observed, and $C = 0$ to denote that a patient was lost to follow-up and the endpoint was unobserved. Furthermore, we perform the conservative imputation mentioned above, such that patients treated with NFX who are lost to follow-up are assumed to not have seroconverted (death and other potential long-term side effects are not a concern for NFX [21]). This allows us to conservatively approximate $E[Y(1)|A=1]$ as $E[Y|A=1]$ – the observed outcome rate among the treated. For approximating $E[Y(0)|A=1]$, if the

variables W contain all common causes of treatment, loss-to-follow-up, and outcome, then it can be proved mathematically that

$$E[Y(0)|A = 1] = E[E(Y|C = 1, A = 0, W)|A = 1],$$

where the right-hand side of the above expression can be estimated by running a regression of the outcome on baseline variables among observed controls and using that regression to predict the outcomes that would have been observed for treated patients had they not been treated. This is accomplished by averaging the predicted outcomes over the empirical distribution of W among the treated. Estimators with better performance are also available, we refer the reader to our companion article published in this edition of the journal for a discussion on optimal estimation. This yields a target statistical estimand equal to

$$\psi = E[Y|A = 1] - E[E(Y|C = 1, A = 0, W)|A = 1],$$

which, contrary to the causal estimand ψ^c, is a quantity that can be estimated from data. The fundamental problem is that the assumptions required for establishing the equality $\psi^c = \psi$, namely that W contains all common causes of treatment, loss-to-follow-up, and outcome, is unlikely to hold in this study. We must therefore study the so-called *causal gap,* defined as the difference between the causal target and the statistical target, i.e., $\psi - \psi^c$. In the supplementary materials, we show that this causal gap may be bounded as $\psi - \psi^c \leq E[Y(0)|A=1]$. The right-hand side of this inequality is precisely the probability of spontaneous seroreversion that we would have observed for treated patients had they not been treated.

Consider now the null hypothesis of no treatment effect of Nifurtimox, i.e., $\psi^c \leq 0$. According to the above discussion, this hypothesis is true if the hypothesis $\psi \leq E[Y(0)|A=1]$ is true. While the hypothesis $\psi^c \leq 0$ cannot generally be tested, the hypothesis $\psi \leq E[Y(0)|A=1]$ can be tested for varying user-given conjectured levels of the probability of spontaneous seroreversion. If this hypothesis is rejected even for the largest feasible values of the probability of spontaneous seroreversion, then we can be confident that the causal hypothesis of no treatment of NFX may also be rejected.

The probability of spontaneous seroreversion is a *sensitivity parameter,* meaning it is a parameter that is useful for sensitivity analyses. We do not know its true value, but we can make conjectures about plausible values based on our knowledge of the subject matter. It is important that pre-alignment and prespecification of the range of plausible values occurs prior to the conduct of the analyses, in order to avoid researcher degrees of freedom [3] and other possible biases.

Results of sensitivity analysis for the effect of Nifurtimox on the Chagas disease

We analyzed the data of Fabbro et al. using the above sensitivity analysis. The statistical significance of the hypothesis test is given in Figure 1 as a function of conjectured values for the probability of spontaneous seroreversion. This figure allows us to conclude that, if we believe that the probability of spontaneous seroreversion among the treated is smaller than 0.19, we can reject the hypothesis of no treatment effect of Nifurtimox with a with a two-sided type I error rate of at most 0.05. All the epidemiologic studies as well as biological knowledge about the Chagas disease suggest that the rate of spontaneous seroreversion is smaller than 5%.
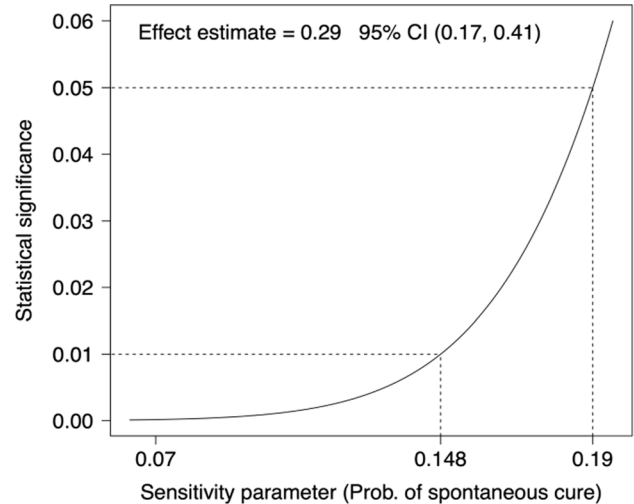
Figure 1. Sensitivity analysis for the effect of Nifurtimox in the treatment of the Chagas disease.

Additional technical details about this sensitivity analysis as well as the methods used to estimate the causal parameter ψ are available in the supplementary materials.

Current landscape and existing methods for sensitivity analysis

In this section, we review some of the most common methods for sensitivity analysis and provide comments on their strengths and weaknesses. This review is not exhaustive, and the reader is referred to Liu *et al.* [22] and Richardson *et al.* [23] for more extensive reviews.

Semiparametric sensitivity analysis

The assumption of no unmeasured confounders may be stated mathematically in multiple ways. One of them is the assumption of independence between the potential outcomes $Y(a)$ and the exposure A of interest (often conditional on observed confounders W). The main idea behind semiparametric sensitivity analyses is to posit a model relating to the potential outcomes to the exposure of interest [24–26]. For instance, one may posit that the probability of exposure $A = 1$ conditional on potential outcome $Y(a)$ (and possibly covariates W) follows a main-terms logistic regression model. The causal effect of A on Y is then identifiable except for the coefficient in front of $Y(a)$ in the above logistic regression. This coefficient, interpreted as the log-odds ratio between $Y(a)$ and A, can be used as a sensitivity parameter that quantifies the magnitude of unmeasured confounding. Analysis may therefore proceed by estimating the causal effect for multiple conjectured values of the sensitivity parameter and judging the plausibility of each such value based on subject-matter expert knowledge.

A disadvantage of this approach is that the sensitivity analysis itself requires positing untestable assumptions about a model relating to the exposure and the potential outcomes. It is unclear whether misspecification of this model carries serious implications in terms of bias, but it would generally be preferable to rely on sensitivity analyses that do not make extra assumptions. Relatedly, the sensitivity parameter must be informed by subject-matter expert knowledge, but it is defined in a scale that is unintelligible and refers to a convenient mathematical construction (e.g., an odds

ratio in a logistic regression between $Y(a)$ and A) rather than a fundamental property of nature. This makes it hard for subject-matter experts to judge on the plausibility of specific values of the sensitivity parameter.

As an example of this approach, Franks *et al.* [26] conduct a sensitivity analysis on the effect of antihypertensives on diastolic blood pressure (DBP) using the National Health and Nutrition Examination Survey data. They conclude that, if one is willing to assume that the adjusted odds of receiving antihypertensives in a logistic regression model increases by 1.01 for every additional mmHg in hypothetical counterfactual DBP outcomes under treatment or control, then an otherwise protective but non-significant effect becomes significant. This example illustrates the difficulty in assessing the plausibility of the sensitivity parameter values. Is a logistic regression adjusted odds ratio between counterfactual DBP outcomes and antihypertensives of 1.01 plausible or implausible? The answer to that question depends non-trivially on the variables included in the model as well as on the correctness of the model, which is potentially as difficult to assess as the original "no unmeasured confounder" assumption.

Nonparametric sensitivity analysis

In contrast to semiparametric sensitivity analyses, nonparametric analyses make no assumptions on the functional form of the relations between variables. This type of sensitivity analysis focuses directly on studying the causal gap with a goal of establishing bounds on it that may be used as sensitivity parameters.

The analysis of the effectiveness of Nifurtimox in the treatment of the Chagas disease presented above is an example of a nonparametric sensitivity analysis. A more general version of this idea has been developed [19,27], where the goal is to construct bounds on the causal gap using sensitivity parameters that have immediate substantive interpretations, so that the plausibility of their values can be easily judged using a-priori subject-matter knowledge (e.g., the probability of spontaneous seroreversion).

A second example of a nonparametric sensitivity analysis uses E-values [28,29] to posit the existence of an unmeasured confounder U and creates bounds on the causal gap in terms of conjectured magnitudes of the $U \rightarrow A$ and $U \rightarrow Y$ relations on a risk ratio scale. These risk ratios are then used as sensitivity parameters. This approach generalizes the sensitivity analysis of Cornfield *et a.l* [5] in the sense that it seeks to find the minimum effect of an unmeasured confounder such that the observed effect would be completely explained away. As an example of the use of E-values, Bosch *et al.* [30] recently studied the effectiveness of fludrocortisone and hydrocortisone on death or discharge to hospice in the treatment of patients with septic shock. Their analyses adjusting for measured confounders found a significant absolute risk difference of −3.7% (95% CI −4.2% – −3.1%) comparing hydrocortisone-fludrocortisone vs hydrocortisone alone. Their sensitivity analysis using E-values concluded that an unmeasured confounder that increases the likelihood of treatment and outcome by 37% would be sufficient to explain away the significant effect found in the analyses.

Importantly, E-values cannot accommodate complex high-dimensional confounders. Furthermore, some E-value analyses make strong assumptions, such as assuming that the risk ratio between the unmeasured confounder and the exposure is equal to the risk ratio between the unmeasured confounder and the

outcome, as well as the assumption that the prevalence of the uncontrolled confounder among the exposed is 100% [31]. Multiple other methods exist that rely on similar ideas but make parametric assumptions on the $U \rightarrow A$ and $U \rightarrow Y$ relations to incorporate complex confounders [32–34], although methods relaxing these assumptions also exist [35–37].

Identification Bounds

Identification bounds are not formally a method for sensitivity analysis in the sense that they do not rely on assessing plausible values for a sensitivity parameter. However, they serve the same purpose of providing information about causal relationships in the presence of unmeasured confounders. The main idea behind identification bounds is to estimate an interval (different from a confidence interval) that bounds the causal effect of a treatment, where this interval is guaranteed to contain the causal effect under no assumptions on the extent of unmeasured confounding.

For example, Bhattacharya *et al.* [38] used identification bounds to study the effect of right heart catheterization (RHC) on 30-day mortality among ICU patients. There is considerable debate in the clinical literature regarding the use of RHC as a diagnostic tool, and its use has been recommended only when there is uncertainty about the best treatment [39]. Therefore, unmeasured confounding is a likely threat to the conclusions of observational analyses of effects of RHC. Using two different types of analyses that allow for any kind of unmeasured confounding, Bhattacharya et al. found that RHC had either a null or a protective effect on 30-day mortality, whereas prior studies that assumed no unmeasured confounders had found RHC to increase 30-day mortality [40]. Although the analyses of Bhattacharya et al. rely on an instrumental variable assumption, multiple identification bounds in the literature do not require this or any other assumption [41].

Identification bounds are most commonly used in the econometrics literature, but they have also been used to assess the comparative effectiveness of treatments in RWD, as illustrated by the above example. Because it relies on few assumptions and has an ambitious goal, this methodology sometimes results in wide bounds that may be uninformative. Manski [41] and Molinari [42] provide a comprehensive review of existing methods for identification bounds.

Negative controls

Additional methods such as negative control treatments and outcomes may be used to rule out the possibility that observed adjusted associations are due to unobserved confounding [43]. For instance, Dickerman *et al.* [44] recently used RWD to assess the comparative effectiveness of COVID-19 vaccines in a real-world population of US veterans. It is thought that COVID-19 vaccines cannot possibly influence infection status in the 10-day period following the vaccination. Thus, infection status at 10 days post-vaccination may be used as a negative control outcome. Specifically, if a procedure purported to estimate causal effects yields a non-null effect on this outcome, then that procedure must be ruled out as giving biased causal estimates. A review by Shi *et al.* [45] provides further examples of successful use of negative controls in applied research. This kind of ad-hoc negative control does not guarantee that an association may be interpreted causally but can be used to rule out non-causal associations, although recent efforts have been made in the statistics literature to formalize the

Table 2. Types of sensitivity analyses described and their advantages and disadvantages

Type of sensitivity analysis	Disadvantages	Advantages	Example
Semiparametric	• Requires arbitrary models for unmeasured variables. • Requires positing plausible values for the unintelligible coefficients of the above model.	• Mathematical convenience.	• Franks *et al.* [26] – antihypertensives and diastolic blood pressure.
Nonparametric	• Typically, none, although some methods such as instances of the E-value might use implausible assumptions [31].	• Requires positing plausible values for intelligible scientific quantities (e.g., spontaneous probability of Chagas seroreversion).	• Nifurtimox on Chagas disease, discussed in this manuscript.
Negative controls	• Does not conclusively guarantee that associations are causal.	• Requires positing an outcome with a null treatment effect, which is often feasible.	• Dickerman *et al.* [44] – COVID vaccines
Identification bounds	• Bounds are often too wide to be informative.	• Operates with few or no assumptions.	• Bhattacharya *et al.* [38] – right heart catheterization and 30-day mortality

use of negative controls for the identification of causal parameters in the presence of unmeasured confounding [46,47].

Table 2 summarizes the assumptions, advantages, and disadvantages of the above types of sensitivity analysis.

Sensitivity analysis considerations when using RWD for regulatory science

Prespecification

Prespecification of a study refers to the publication in complete detail of the study design and analysis plan before all data are collected and analyses are conducted [48]. As with all aspects of a data analysis, sensitivity analyses must be fully prespecified to appropriately control type I error and avoid biases due to researcher degrees of freedom [49]. FDA guidance does allow for choices to be made using blinded data with prespecification of the plan after such examination [50]. Prespecification of the analysis must include the range of plausible values for the sensitivity parameter, which may be based on prior literature or consensus in the substantive field. For instance, a prespecified analysis plan for the case study of the effect of Nifurtimox on the Chagas disease may have conservatively prespecified 10% as the maximum possible rate of spontaneous seroreversion among the treated, based on prior literature that suggests that this rate is of at most 5% [17–19].

The need for prespecification means that it is important that the sensitivity parameters used have an interpretation that corresponds to interpretable phenomena rather than convenient mathematical formalizations, as in our illustration on the effect of NFX on the Chagas disease. This ensures that prespecified plausible values may be obtained through consultation with experts or the literature. The need for prespecification makes it harder to use sensitivity parameters interpreted as the coefficient relating the exposure A and the potential outcome $Y(a)$ in a logistic or linear regression model. Likewise, analyses that rely on a sensitivity parameter interpreted in terms of the strength of the associations $U \to A$ and $U \to Y$ will usually require that the unmeasured confounder U is specified and described in terms of real-world phenomena, even if it is not possible to measure it. Arbitrary unspecified confounders will make it difficult for subject-matter experts to obtain prior information that can inform plausible values for the associations $U \to A$ and $U \to Y$.

Sensitivity analyses with assumptions

Some methods for sensitivity analysis use statistical models to obtain mathematical expressions of violations of the assumptions of the model. For example, a strand of the literature makes the assumption that the probability of treatment A within strata of the potential outcome $Y(a)$ (and possibly measured confounders) follows a logistic regression model [24–26]. Other methods directly assume statistical models that capture the dependence between the outcome Y and a hypothetical unmeasured confounder U, for example assuming that they are linearly related [34]. Like all statistical models, these models are subject to misspecification. For instance, it could be the case that the relation between the confounder U and the outcome Y is quadratic, so that a linear approximation will fail to account for unmeasured confounding. Unlike statistical models applied to real data, models for unobserved variables such as $Y(a)$ and U are not testable. Therefore, while using sensitivity analyses based on models is certainly better than not performing a sensitivity analysis at all, it is preferable to use sensitivity analyses that make no assumptions about the mathematical nature of the unmeasured confounding.

Summary and conclusions

Sensitivity analysis is an important tool that can help researchers test whether causal conclusions obtained from analyses of observational data are robust to violations of assumptions of the causal model. The routine use of sensitivity analyses with RWD increases the trustworthiness of effectiveness conclusions for regulatory science. Sensitivity analyses are most likely to be useful and informative when other aspects of the study (described in our companion paper on the *causal roadmap*) are also carefully designed. That is, sensitivity analyses on their own are not a panacea and cannot save a poorly designed and conducted analysis of RWD. As with other study aspects, prespecification of sensitivity analyses for RWD in regulatory settings is crucial to avoid wrong conclusions due to researcher degrees of freedom [51]. Most sensitivity analyses require auxiliary scientific information (e.g., the probability of spontaneous seroreversion in the Chagas disease example discussed) to produce meaningful conclusions, although some methods such as those based on identification bounds can sometimes produce meaningful conclusions without such knowledge.

In this paper, we focused on an illustration of sensitivity analyses for the assumption of unmeasured confounding, but causal models often entail other important assumptions which may also be subject to sensitivity analysis.

There is a vast literature on sensitivity analysis for causal inference with many fields contributing distinct approaches and tools. For instance, the computer science and machine learning community has developed software tools such as PyWhy [52] that help scientists capture causal assumptions and apply sensitivity analyses and other refutations. Furthermore, there are numerous developed and emerging methods that rely on different assumptions appropriate to a variety of scenarios, such as the identification of secondary small-scale or continuous experiments to infer or validate causal assumptions (i.e., adaptive, active sampling, or reinforcement learning) [53,54] and explorations of large language models as a source of domain knowledge for semi-automated critiquing and refinement of researchers' causal assumptions [55]. Such tools and emerging methods and their requirements should be considered and assessed carefully before use in regulatory science. In all cases, it is important to specify sensitivity analysis that have at least two important properties: (i) their conclusions do not rely on further untestable assumptions, and (ii) the sensitivity parameter has a clear scientific interpretation so that prespecification of a plausible range of values is possible from available subject-matter knowledge.

References

1. **Gabay M.** 21st century cures act. *Hosp Pharm.* 2017;**52**(4):264–265. doi: 10.1310/hpx5204-264.

2. **Dang LE, Gruber S, Lee H**, *et al.* A Causal Roadmap for generating high-quality real-world evidence. *arXiv [statME].* http://arxiv.org/abs/2305.06850. Accessed May 11, 2023.

3. **Guideline IHT.** Statistical principles for clinical trials. International conference on harmonisation E9 expert working group. *Stat Med.* 1999;**18**:1905–1942.

4. **Permutt T.** Sensitivity analysis for missing data in regulatory submissions. *Stat Med.* 2016;**35**(17):2876–2879.

5. **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**:173–203.

6. **Hall W.** Cigarette century: the rise, fall and deadly persistence of the product that defined America. *Tob. Control.* 2007;**16**(5):360.

7. **Rosenbaum PR.** Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika.* 1987;**74**(1):13–26.

8. **Grodstein F, Manson JE, Stampfer MJ.** Postmenopausal hormone use and secondary prevention of coronary events in the nurses' health study. a prospective, observational study. *Ann Intern Med.* 2001;**135**(1):1–8.

9. **Stampfer MJ, Colditz GA.** Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev Med.* 1991;**20**(1):47–63.

10. **Rossouw JE, Anderson GL, Prentice RL**, *et al.* Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *JAMA.* 2002;**288**(3):321–333.

11. **Yu R, Small DS, Rosenbaum PR.** The information in covariate imbalance in studies of hormone replacement therapy. *Ann Appl Stat.* 2021;**15**(4):2023–2042. doi: 10.1214/21-aoas1448.

12. **Cagnacci A, Venier M.** The controversial history of hormone replacement therapy. *Medicina (Kaunas).* 2019;**55**(9):602.

13. **Shikanai-Yasuda MA, de Almeida EA, López MC, Delgado MJP.** Chagas disease: a parasitic infection in an immunosuppressed host. *Chagas Disease: A Neglected Tropical Disease.* 2020: 213–234.

14. **Hochberg NS, Montgomery SP.** Chagas disease. *Ann Intern Med.* 2023;**176**(2):ITC17–ITC32.

15. **Altcheh J, Castro L, Dib JC**, *et al.* Prospective, historically controlled study to evaluate the efficacy and safety of a new paediatric formulation of nifurtimox in children aged 0 to 17 years with Chagas disease one year after treatment (CHICO). *PLoS Negl Trop Dis.* 2021;**15**(1):e0008912.

16. **Fabbro DL, Streiger ML, Arias ED, Bizai ML, del Barco M, Amicone NA.** Trypanocide treatment among adults with chronic Chagas disease living in Santa Fe city (Argentina), over a mean follow-up of 21 years: parasitological, serological and clinical evolution. *Rev Soc Bras Med Trop.* 2007;**40**(1):1–10. doi: 10.1590/s0037-86822007000100001.

17. **Sosa Estani S, Segura EL, Ruiz AM, Velazquez E, Porcel BM, Yampotis C.** Efficacy of chemotherapy with benznidazole in children in the indeterminate phase of Chagas' disease. *Am J Trop Med Hyg.* 1998;**59**(4):526–529.

18. **Villar JC, Herrera VM, Carreño JGP**, *et al.* Correction to: nifurtimox versus benznidazole or placebo for asymptomatic Trypanosoma cruzi infection (Equivalence of usual interventions for trypanosomiasis - EQUITY): study protocol for a randomised controlled trial. *Trials.* 2019;**20**(1). doi: 10.1186/s13063-019-3630-y.

19. **Díaz I, van der Laan MJ.** Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *Int J Biostat.* 2013;**9**(2):149–160.

20. Use of Electronic Health Record Data in Clinical Investigations Guidance for Industry. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-electronic-health-record-data-clinical-investigations-guidance-industry

21. **Castro JA, de Mecca MM, Bartel LC.** Toxic side effects of drugs used to treat Chagas' disease (American trypanosomiasis). *Hum Exp Toxicol.* 2006;**25**(8):471–479.

22. **Liu W, Kuramoto SJ, Stuart EA.** An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prev Sci.* 2013;**14**(6):570–580.

23. **Richardson A, Hudgens MG, Gilbert PB, Fine JP.** Nonparametric bounds and sensitivity analysis of treatment effects. *Stat Sci.* 2014;**29**(4):596–618.

24. **Scharfstein DO, Rotnitzky A, Robins JM.** Adjusting for nonignorable drop-out using semiparametric nonresponse models: rejoinder. *J Am Stat Assoc.* 1999;**94**(448):1135. doi: 10.2307/2669930.

25. **Robins JM, Rotnitzky A, Scharfstein DO.** Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: *Statistical Models in Epidemiology, the Environment, and Clinical Trials.* New York, NY: Springer New York, 2000, 1–94.

26. **Franks A, D'Amour A, Feller A.** Flexible sensitivity analysis for observational studies without observable implications. *J Am Stat Assoc.* 2020;**115**(532):1730–1746. doi: 10.1080/01621459.2019.1604369.

27. **Díaz I, Luedtke AR, van der Laan MJ.** Sensitivity Analysis. In: *Targeted Learning in Data Science.* Cham: Springer Series in Statistics, Springer; 2018. doi: 10.1007/978-3-319-65304-4_27

28. **VanderWeele TJ, Ding P.** Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med.* 2017;**167**(4):268–274.

29. **Ding P, VanderWeele TJ.** Sensitivity analysis without assumptions. *Epidemiology.* 2016;**27**(3):368–377. doi: 10.1097/ede.0000000000000457.

30. **Bosch NA, Teja B, Law AC, Pang B, Jafarzadeh SR, Walkey AJ.** Comparative effectiveness of fludrocortisone and hydrocortisone vs hydrocortisone alone among patients with septic shock. *JAMA Intern Med.* 2023;**183**(5):451–459.

31. **MacLehose RF, Ahern TP, Lash TL, Poole C, Greenland S.** The importance of making assumptions in bias analysis. *Epidemiology.* 2021; **32**(5):617–624.

32. **Zhang B, Tchetgen Tchetgen EJ.** A semi-parametric approach to model-based sensitivity analysis in observational studies. *J R Stat Soc Ser A Stat Soc.* 2022;**185**(S2):S668–S691. doi: 10.1111/rssa.12946.

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

34. **Imbens GW.** Sensitivity to exogeneity assumptions in program evaluation. *Am Econ Rev.* 2003;**93**(2):126–132. doi: 10.1257/000282803321946921.

35. **Yadlowsky S, Namkoong H, Basu S, Duchi J, Tian L.** Bounds on the conditional and average treatment effect with unobserved confounding factors. *Ann Stat.* 2022;**50**(5). doi: 10.1214/22-aos2195.

36. **Saltelli A, Tarantola S, Campolongo F, Ratto M.** *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*, vol. 1. New York: John Wiley & Sons; 2004.

37. **Shen C, Li X, Li L, Were MC.** Sensitivity analysis for causal inference using inverse probability weighting. *Biom J.* 2011;**53**(5):822–837. doi: 10.1002/bimj.201100042.

38. **Bhattacharya J, Shaikh AM, Vytlacil E.** Treatment effect bounds: an application to Swan-Ganz catheterization. *J Econom.* 2012;**168**(2):223–243.

39. **Kittleson MM, Prestinenzi P, Potena L.** Right heart catheterization in patients with advanced heart failure: when to perform? How to interpret? *Heart Fail Clin.* 2021;**17**(4):647–660.

40. **Connors AF Jr, Speroff T, Dawson NV**, *et al.* The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT investigators. *JAMA.* 1996;**276**(11):889–897.

41. **Manski CF.** *Partial Identification of Probability Distributions.* New York, NY: Springer Science & Business Media; 2003.

42. **Molinari F.** *Microeconometrics with partial identification.* Handbook of Econometrics; 2020:355–486. https://www.sciencedirect.com/science/article/abs/pii/S1573441220300027?via%3Dihub

43. **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.

44. **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.

45. **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.

46. **Shi X, Miao W, Nelson JC, Tchetgen EJT.** Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *J R Stat Soc Series B Stat Methodol.* 2020;**82**(2):521–540.

47. **Miao W, Geng Z, Tchetgen Tchetgen E.** Identifying Causal effects with proxy variables of an unmeasured confounder. *Biometrika.* 2018;**105**(4): 987–993.

48. **Williams RJ, Tse T, Harlan WR, Zarin DA.** Registration of observational studies: is it time? *CMAJ.* 2010;**182**(15):1638–1642.

49. **Kasy M, Spiess J.** *Rationalizing Pre-Analysis Plans: Statistical Decisions Subject to Implementability*, No. 975. University of Oxford, Department of Economics, 2022.

50. **Adaptive designs for clinical trials of drugs and biologics**. Published 2019. https://www.fda.gov/media/78495/download. Accessed March 21, 2023.

51. **Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM, van Assen MALM.** Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Front Psychol.* 2016;**7**:1832.

52. **Sharma A, Kiciman E**. DoWhy: An end-to-end library for causal inference. *arXiv [statME].* http://arxiv.org/abs/2011.04216. Accessed November 9, 2020.

53. **Zhu S, Ng I, Chen Z**. Causal discovery with Reinforcement Learning. *arXiv [csLG].* http://arxiv.org/abs/1906.04477. Accessed June 11, 2019.

54. **Gao XE, Hu JG, Chen B, Wang YM, Zhou SB.** Causal discovery approach with reinforcement learning for risk factors of type II diabetes mellitus. *BMC Bioinform.* 2023;**24**(1):296.

55. **Kıcıman E, Ness R, Sharma A, Tan C**. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv [csAI].* http://arxiv.org/abs/2305.00050. Accessed April 28, 2023.

56. **Gabay M.** 21st century cures act. *Hosp Pharm.* 2017;**52**(4):264–265. doi: 10.1310/hpx5204-264.