Received: 27 December 2019 | Revised: 17 August 2020 | Accepted: 20 August 2020

DOI: 10.1002/pds.5117

WILEY

REVIEW

Assessing the impact of unmeasured confounders for credible and reliable real-world evidence

Xiang Zhang[1] | James D. Stamey[2] | Maya B. Mathur[3]

[1]Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana

[2]Department of Statistics, Baylor University, Waco, Texas

[3]Quantitative Sciences Unit, Stanford University, Stanford, California

Correspondence
James D. Stamey, Baylor University, Waco, TX.
Email: james_stamey@baylor.edu

Abstract

Purpose: We review statistical methods for assessing the possible impact of bias due to unmeasured confounding in real world data analysis and provide detailed recommendations for choosing among the methods.

Methods: By updating an earlier systematic review, we summarize modern statistical best practices for evaluating and correcting for potential bias due to unmeasured confounding in estimating causal treatment effect from non-interventional studies.

Results: We suggest a hierarchical structure for assessing unmeasured confounding. First, for initial sensitivity analyses, we strongly recommend applying a recently developed method, the E-value, that is straightforward to apply and does not require prior knowledge or assumptions about the unmeasured confounder(s). When some such knowledge is available, the E-value could be supplemented by the rule-out or array method at this step. If these initial analyses suggest results may not be robust to unmeasured confounding, subsequent analyses could be conducted using more specialized statistical methods, which we categorize based on whether they require access to external data on the suspected unmeasured confounder(s), internal data, or no data. Other factors for choosing the subsequent sensitivity analysis methods are also introduced and discussed, including the types of unmeasured confounders and whether the subsequent sensitivity analysis is intended to provide a corrected causal treatment effect.

Conclusion: Various analytical methods have been proposed to address unmeasured confounding, but little research has discussed a structured approach to select appropriate methods in practice. In providing practical suggestions for choosing appropriate initial and, potentially, more specialized subsequent sensitivity analyses, we hope to facilitate the widespread reporting of such sensitivity analyses in non-interventional studies. The suggested approach also has the potential to inform pre-specification of sensitivity analyses before executing the analysis, and therefore increase the transparency and limit selective study reporting.

KEYWORDS

causal inference, pharmacoepidemiology, practical recommendation, real world evidence, sensitivity analyses, unmeasured confounding

> "Real-world evidence can only correct for biases that researchers already understand. By randomly assigning patients to one treatment or another, clinical trials rely on chance to cancel out any biases, whether researchers are aware of them or not."

1 | INTRODUCTION

In July 2019, *STAT* published an article[1] "Attempt to replicate clinical trials with real-world data generates real-world criticism, too" and the criticism above generated our immediate attention. Is it a legitmate criticism? Of course it is. As researchers who are dealing with non-interventional studies and associated causal inference problems on a daily basis, we are certainly aware of the challenges caused by the lack of randomization and among them unmeasured confounding is an obvious and important one. However, the existence of these challenges does not necessarily means RWE can only correct for biases that the researchers already understand. Various methods, both at the design and analysis stages of non-interventional studies, have been developed and applied to assess and/or adjust the impact of bias coming from confounders that were not anticipated or not available in the databases analyzed.

Evidence from randomized clinical trials (RCTs) has been considered the gold standard in assessing causal treatment effects of an investigational therapy and will probably remain the cornerstone for that purpose. However, RCTs are not always feasible due to factors such as ethical concerns. For instance, in life-threatening rare-disease areas where there is substantial unmet medical need, it may be considered unethical to randomize patients to a placebo or standard-of-care group, and investigators may instead allocate the investigational drug to all enrolled patients and compare them to external, historical controls. On the other hand, RWE is increasingly becoming a vital component of decision making in health care for patients, clinicians, sponsors, and even regulatory agencies. In the past, regulators used RWE primarily to monitor and evaluate the safety of drug products after they were approved, and only under limited circumstances have considered RWE to support a claim of efficacy.[2] However, with rising awareness and interest in RWE worldwide, an important opportunity has emerged to expand the established paradigm of evidence generation and to harness the potential of RWE for regulatory decisions more often and more broadly. For example, the European Medicines Agency (EMA) considered RWE as a part of adaptive licensing in 2016[3]; The US Food and Drug Administration (FDA) published its framework for an RWE program in 2018[4]; the National Medical Products Administration (NMPA) in China published its draft guidance "Key Considerations in Using Real-World Evidence to Support Drug Development" in 2019.[5] To achieve the ultimate goal of wider acceptance and more responsible utilization of RWE, we have to properly address the criticisms similar to these, so that we can build more credible and reliable RWE. A well-designed non-interventional study based on high-quality data sources may provide evidence that is as credible as that provided by a RCT. A well-known example is the use of hormone replacement therapy (HRT) for post-menopausal women. HRT had previously been supported by positive results from a high-quality and long-running non-interventional study, which suggested that HRT would reduce the risk of heart disease.[6] However, results from a subsequent RCT showed increased cardiovascular risks.[7] Initially, these differences were thought to indicate the weakness of non-interventional studies, but further analyzes determined that both studies had valid results for their patient populations and that discrepancies were probably due to the timing of initiation of hormone therapy in relation to the onset of menopause.[8,9] If this is true, then the RCT and non-interventional study actually showed similar findings. However, because RCT was considered as the ideal methodology to assess causal inference and non-interventional studies was considered as weaker evidence due to lack of randomization, the negative result of this RCT led to widespread abandonment of the HRT therapy, which might have been a mistake.

Pharmacoepidemiology is a research field that includes analysis of routinely collected electronic health data, and has developed good practices on the conduct and report of RWD-based studies. This field also provides the use of its research for regulators' decision, for example, FDA has long performed safety monitoring studies through pharmocoepidemiological research projects under the Sentinel Initiative.[10] Most recently, as part of its RWE program, FDA funded the RCT DUPLICATE project, which intends to replicate the results of 30 RCTs using real-world claims databases.[11] The project aims to see whether

KEY POINTS

- Real world evidence (RWE) is now widely utilized for decision-making in routine clinical practice, and regulators have started to consider evidence from studies based on real-world data (RWD) for their decisions. However, challenges remain on the use of RWE, and a critical one is reducing bias in causal treatment effect estimates when randomization is not available or feasible.
- Controlling confounding and other biases inherent in non-interventional RWD-based studies has been an important task to generate less biased causal evidence and therefore provide more credible and reliable RWE. However, bias due to unmeasured confounding still remains one of the major criticisms of the credibility of RWE.
- Various analytical methods have been proposed to address unmeasured confounding, but little research has discussed a structured approach to select appropriate methods in practice. In this paper, we will discuss and provide practical suggestions via a flowchart to illustrate a decision process that researchers could consider in their own studies to mitigate the impact of unmeasured confounding. The suggested approach has the potential to help pre-specify the sensitivity analyses in statistical analysis plans (SAP) for a given non-interventional study.

the claims databases are able to replicate the trial results, and it will assist in the FDA's evaluation of the use of RWE to support new indications for approved drugs or to satisfy post-approval study requirements. Given the close connection between pharmacoepidemiologic research and the RWE generation, properly addressing bias due to confounding in causal inference assessment is an important task for this field. Thus, evaluating and correcting bias due to unmeasured confounding should then be an important part of this task.

2 | CONFOUNDING AND THE "NO UNMEASURED CONFOUNDING" ASSUMPTION

So what is a confounder? Under the potential outcomes framework, the individual-level causal treatment effect is:

$$Y(T=1)-Y(T=0)$$

where $T = 1$ and $T = 0$ represent the two treatments being compared and $Y(T = 1)$ and $Y(T = 0)$ represent the potential outcomes of a subject had he/she received the corresponding treatment. Because an individual can never simultaneously receive $T = 0$ and $T = 1$, these individual-level potential outcomes cannot be observed simultaneously. Accordingly, the potential outcome of the unrealized treatment is often referred to as the "counterfactual" outcome. A key goal of causal inference is to estimate the average treatment effect. In a randomized trial, subjects' potential outcomes are independent of the treatments they actually received, which essentially means that the subjects in each treatment group are comparable to one another, and this "exchangeability" means that the average treatment effect can be estimated without bias. In contrast, in an observational study, subjects in each treatment group may not be comparable if there are variables that affect subjects' probability of receiving the treatment and also independently affect their probability of experiencing the outcome. These variables are called "confounders". If all confounders are measured and adjusted in analysis (eg, via multivariable regression), then the potential outcomes of the subjects in each treatment group (conditional on the confounders) are once again exchangeable, allowing unbiased estimation of the treatment effect. The assumption that all confounders have been measured and adjusted in analysis is sometimes called the "no unmeasured confounding assumption" (NUCA).[12,13]

However, in pharmacoepidemiologic research, it is rarely the case that every possible confounder has been measured and adjusted in the analysis. For instance, RWD are always collected for reasons other than research and they may have missing or limited information on potential confounders such as smoking, body mass index, and disease severity measurements. When estimating a causal treatment effect, ignoring the influence of unmeasured confounders could lead to substantial bias - potentially even reversing the estimated effect. Such reversal can occur, for example, when there is "confounding by indication", in which the medical indication to receive a given treatment is itself a confounder. For example, subjects with greater disease severity may be medically indicated to receive a particular drug, but may also have a greater risk of negative health outcomes, creating confounding that could weaken or even reverse a potentially beneficial effect of the drug.[14] Another common unmeasured confounder in epidemiological studies is healthy user bias. As pointed out by Shrank et al,[15] the healthy user effect is best described as the tendency of patients who receive a preventive therapy to also seek other preventive services or partake in other healthy behaviors. Patients who choose to receive preventive therapy might exercise more, eat a healthier diet, and avoid unhealthy behaviors such as smoking and alcohol use. As a result, a non-interventional study evaluating the effect of a preventive therapy (eg, statin therapy) on a related outcome (eg, myocardial infarction) without adjusting for other related preventive behaviors (eg, healthy diet or exercise) will tend to overstate the effect of the preventive therapy under study.[16]

We refer to such confounders that are not measured for all subjects in the study as "unmeasured confounders" (though, as noted below, they may be measured for a subset of subjects). Given the ubiquity of unmeasured confounding in analyses of RWD, we recommend always conducting sensitivity analyses to assess and/or adjust for its potential impact on estimates of causal effects. Sensitivity analyses assess the extent to which the results of an analysis (eg, the estimated average treatment effect) might change if NUCA does not hold exactly due to the presence of unmeasured confounding. This is critical because, except in the context of a large randomized trial, one can never be certain that NUCA holds. That is, when analyzing RWD, one typically conducts analyses that inherently *assume* that NUCA holds, and sensitivity analyses can help characterize how much the results might change if NUCA is violated to a greater or lesser extent.

In Schenessweiss's 2006 paper,[17] he categorized the unmeasured confounders into two types: one type is "measurable", which means we know those factors are confounding variables but the database we intend to use does not collect the variables or does not completely collect the variables. Another type is variables that are not "measurable" even in principle or at least are measurable only by proxy, and one popular example of this type is the healthy user bias that we mentioned before.

In a particular study utilizing real world data, we distinguish three possible scenarios regarding available information on unmeasured confounders. Firstly, there may be no additional data available, which means the researchers do not know any extra information about the unmeasured confounder. Secondly, there may be internal data available, which means the database does not have the information of confounders on every subject but does for a subsampling of the study population. Thirdly, there may be external data available, which means the database does not have the information of the confounders on any subject but such information are available in patient-level data or summary data external to the study database.

3 | A FRAMEWORK FOR CHOOSING SENSITIVITY ANALYSES FOR UNMEASURED CONFOUNDING ASSESSMENT

In *Pharmacoepidemiology*, Lash and his colleagues published extensively on quantitative bias analysis.[18,19] The quantitative bias analysis

they discussed targets not only the issue of unmeasured confounding but also other forms of internal bias, such as selection bias or misclassification of a covariate in a database. For unmeasured confounding specifically, an early review paper by McMahon[20] discussed the approaches to combat confounding by indication in observational studies, including design approaches such as restriction of study subjects. Later, Schneeweiss outlined strategies to mitigate bias due to unmeasured confounding in general.[17] Because pharmacoepidemiological studies using administrative data are often criticized for their limited ability to collect clinically important confounders, he proposed a framework to adjust residual bias due to unmeasured confounders in those studies. The framework includes several types of methods to address residual bias due to unmeasured confounders: two-stage sampling, external adjustments, design methods such as cross-over designs and active comparator and analysis methods such as instrumental variables or other sensitivity analysis.

Despite the availability of various methods[21-23] and the nice framework laid out by Schnessweiss,[17] in practice it is still difficult to choose the appropriate methods for each individual study, given the huge range of possible confounding factors. In addition, little discussion has been offered on the general problem of data-dredging[24] with respect to those methods for mitigating residual bias due to unmeasured confounding. The International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) issued guidance on the statistical principles that need to be implemented in clinical trials[25] and pre-specification of the analysis is part of those important principles, to prevent the potential bias due to data dredging. However, the principle of pre-specifying analysis has not been widely adopted in observational studies utilizing real world data. The lack of adoption of this principle does not mean that data-dredging is not a concern in observational studies, actually the widespread perception "observational studies propose, RCTs dispose" could partially attribute to the data dredging. For instance, in an observational study using a large real-world database, a large number of associations could be found, when in reality only a few are true associations. Those false positive associations are likely the products of data dredging. Like randomized clinical trials, the validity of conclusions of observational studies is likely to improve by use of a pre-specified analysis plan, so that the pre-planned analyses are distinguished from ad-hoc, data-driven analyses, and journals now encourage researchers to preregister observational studies' protocols and statistical analysis plans.[26,27] Therefore, we would like to provide some suggestions to tackle the practical challenge in assessing the impact of unmeasured confounders. In particular, we think these suggestions could be of great help in statistical analysis pre-specification, thereby improving the transparency and reproducibility of the study results, and ultimately improving the credibility of the RWE.

We summarize those recommendations in the flowchart (Figure 1) below as one possible structured approach to assess/adjust and report the impact of unmeasured confounders on the observed treatment-outcome relationship. This flowchart is an updated version from a previous review paper[23] written by some of the current authors; the updated flowchart additionally incorporates some newly developed methods (eg, E-value[28]) and feedback received from other researchers in the relevant fields.

First, some methods based on quasi-experiments, such as instrumental variables and regression discontinuity designs, do not require the NUCA with respect to the confounders between the treatment and outcome. Critically, though, these quasi-experimental methods do require other assumptions as alternatives to NUCA (such as the availability of an exogeneous instrumental variable), and sometimes along with additional alternative assumptions. Thus, it is important to carefully assess potential violations of these alternative assumptions. For details about how to use Instrumental variables/regression discontinuity for causal inference, please refer to Angrist et al,[29] Baiocchi et al,[30,31] Brookhart et al,[32] Mckenzie et al,[33] Cook,[34] Imbens and Lemieux,[35] and the references therein.

If researchers instead apply methods that do need NUCA for the treatment-outcome relationship (eg, regression, propensity score methods, G-estimation, etc.), but are concerned that the assumption may be violated (as is essentially always the case with RWD, as described above), then we suggest conducting sensitivity analyses to quantitatively assess the robustness of the treatment-outcome association to potential unmeasured confounding. Depending on the complexity of the implementation and the assumptions needed, we further categorize sensitivity analyses as "initial" or "subsequent" sensitivity analyses. The main purpose of these initial sensitivity analyses is to test the robustness of the study findings to potential unmeasured confounding in a straightforward manner that requires few assumptions, is applicable for both "measurable" and "unmeasurable" types of unmeasured confounders, and is easy to interpret. Given those considerations, the E-value,[28] the array approach,[17] or the rule-out method[17] are the ones we propose as "initial" sensitivity analysis.

The E-value is a recently proposed approach[28] that is fairly easy to apply in practice and makes minimal assumptions regarding the structure of unmeasured confounding, as described below. The E-value is defined as the minimum strength of association on the risk ratio scale that unmeasured confounder(s) would need to have with both the treatment and the outcome, conditional on the measured covariates, to fully "explain away" the observed treatment-outcome association in the sense that the observed association is compatible with a truly null causal effect (or, alternatively, is compatible with a causal effect of a specific value). A large E-value indicates that the observed association is relatively robust to unmeasured confounding, because it would take considerable unmeasured confounding to explain away the observed association. In contrast, a small E-value indicates that the observed association is relatively sensitive to unmeasured confounding, because relatively weak unmeasured confounding could potentially explain away the effect. In practice, the E-value can be calculated as a simple non-linear transformation of the observed relative risk, *RR*, and is given by E-value = $RR + \sqrt{RR(RR-1)}$. The same formula can be applied to effect measures other than relative risks by using various effect-size transformations.[28] A website (evalue-calculator.com) and R package (EValue) are available to calculate E-values for a variety of effect measures.[36]

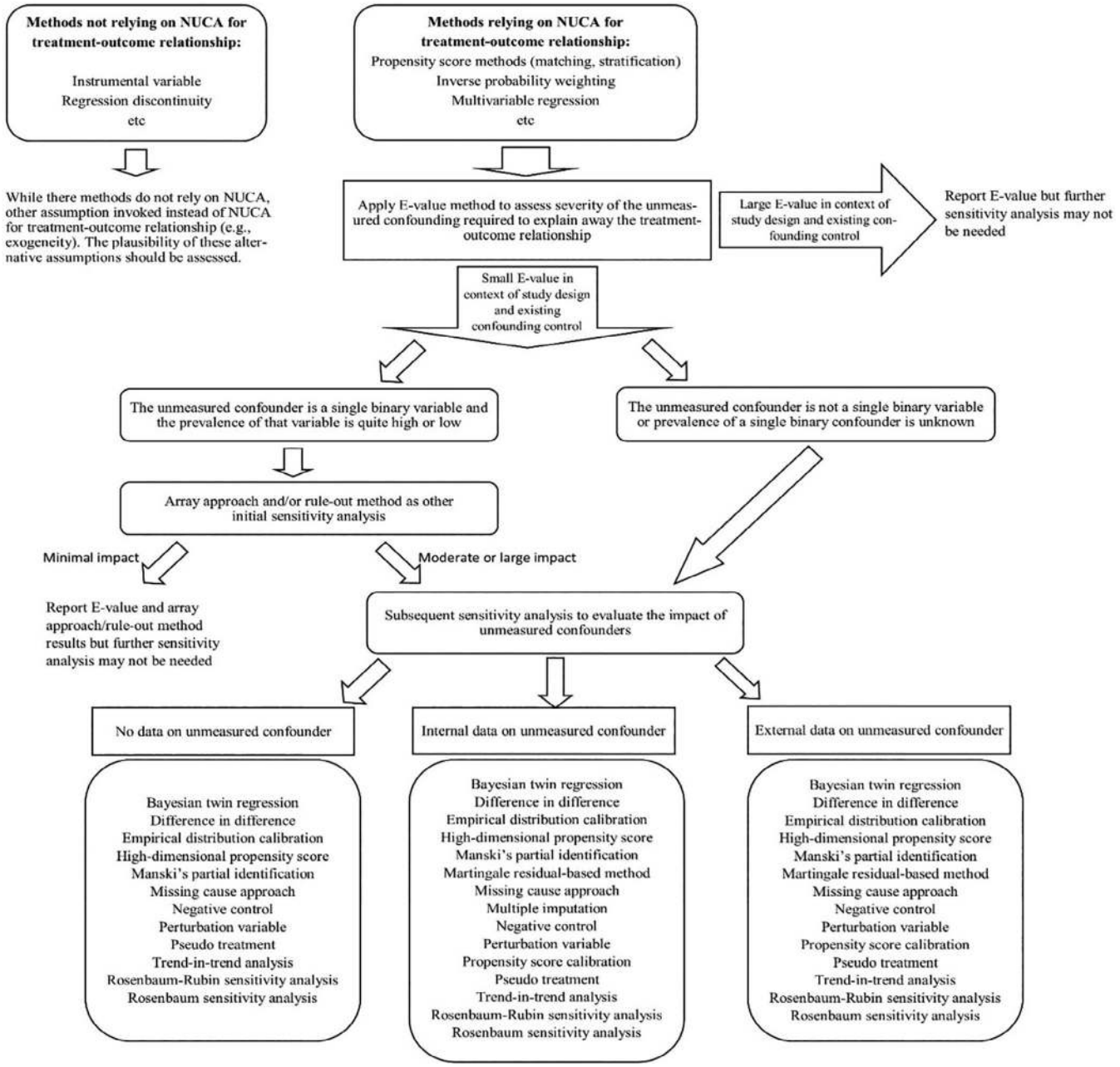

FIGURE 1 Suggested steps to evaulate the impact of unmeasured confounders. "NUCA" = "no unmeasured confounding assumption"

A strength of the E-value approach is that it makes minimal assumptions regarding the structure of unmeasured confounding (eg, it does not assume the unmeasured confounder is binary), nor does it assume that there is a single unmeasured confounder. When there are multiple unmeasured confounders, the minimum strengths of associations between the unmeasured confounders and the treatment (or outcome) are interpreted in terms of the strongest association that is produced by comparing any two categories of the entire vector of unmeasured confounders.[37] The E-value also does not require assumptions regarding the prevalence or distribution of the unmeasured confounder(s), which is both a strength and a limitation. That is, it is *possible* that if there existed unmeasured confounder(s) with confounding strengths at least as large as the E-value, such confounder(s) could potentially explain away the effect, but not *all* confounders with confounding strengths at least as large as the E-value are capable of explaining away the effect.[28]

Like all sensitivity analyses, the E-value should be reported and interpreted thoughtfully.[38] In particular, its interpretation depends on context, particularly on the measured confounders that have been adjusted in analysis.[28] Suppose two studies both had an E-value of 2.5, but one study controlled for a broad range of potential confounders while the other did not control for any confounders. Then the first study could be considered more robust to unmeasured confounding because in that study, unmeasured confounder(s) would

have to be associated with both the treatment and the outcome by relative risks of 2.5 each, above and beyond all of the confounders that have already been measured and adjusted in the analysis. Further guidance and recommendations on the practical interpretation of the E-value, and discussion of its limitations, is provided elsewhere.[38]

We suggest reporting the E-value for all non-interventional studies in which the objective is causal inference and in which the proposed methods to control confounding bias rely on NUCA. As described above, the E-value is a conservative measure in that, rather than making assumptions about the distribution or type of unmeasured confounder(s), it considers the hypothetical worst-case effects of unmeasured confounder(s) with specific strengths of association with the treatment and outcome.[28] This is a strength of the measure, in that it can be applied even when no information is available about the unmeasured confounder(s), but is also a limitation, in that when some information is indeed available, the E-value may be quite conservative.[28] In cases in which researchers have substantive knowledge that there is a specific unmeasured confounder, and its prevalence is known to be quite low or quite high, the array approach and rule-out methods, described below, may be less conservative than the E-value because they incorporate these assumptions regarding the prevalence. For instance, if a binary unmeasured confounder is strongly associated with the treatment and outcome but its prevalence is very low, then the confounder may produce only relatively little confounding bias in practice, but the E-value may be very small. Thus, if investigators have confident substantive knowledge regarding which specific unmeasured confounder(s) are of concern and of their prevalence, other approaches to sensitivity analysis incorporating these additional assumptions, such as the rule-out and array methods described below, may be preferable, especially if the E-value method results in a small E-value. The array approach[17] is applicable if there is a single binary unmeasured confounder. It expresses the true relative risk (RR) as a function of observed relative risk (ARR), the prevalence of the binary confounder in the treatment (P_{c1}) and comparison group (P_{c0}), and the strength of association between the binary confounder and the outcome (RR_{CD}). If plausible value combinations of (P_{c1}, P_{c0}, RR_{CD}) were able to move ARR to 1, then the impact of the unmeasured confounder is not ignorable and subsequent sensitivity analysis is needed. The rule-out method[17] also applies to the scenario when a single, binary unmeasured confounder is present. Under the hypothesis that there is no treatment-outcome relationship, ARR could be written as a function of the prevalence of the binary confounder in overall population (P_c) and treated subjects (P_{c1}), the prevalence of the treated subjects (P_E), and the strength of association between the binary confounder and the outcome (RR_{CD}). Given ARR, P_c, and P_E, if plausible value combinations of (P_{c1}, RR_{CD}) were able to move the ARR to 1, then the impact of unmeasured confounders is not ignorable and subsequent sensitivity analysis is needed. Back to the example above, while the E-value may be small, the array approach and/or rule-out method could suggest that such an unmeasured confounder would not greatly influence the estimated treatment effect, given its low prevalence.

If these initial sensitivity analyses using the E-value, the rule-out method, and/or the array method suggest that implausibly strong unmeasured confounding would be required to explain away or meaningfully reduce the observed treatment-outcome association, we recommend reporting these sensitivity analysis results, and subsequent sensitivity analyses may not be needed. On the other hand, if these initial sensitivity analyses suggest that relatively weak unmeasured confounding might be able to explain away or meaningfully reduce the observed association, then we would suggest applying subsequent, more detailed sensitivity analyses and reporting these results along with the results of initial sensitivity analyses. To date, various analytical methods could be used for such subsequent sensitivity analysis. These methods include, but are not limited to: Bayesian twin regression modeling,[39,40] difference in difference,[41,42] empirical distribution calibration,[43,44] high-dimensional propensity score,[45] Manski's partial identification,[46] martingale residual-based method,[47,48] missing cause approach,[49] multiple imputation,[50,51] negative control,[52,53] perturbation variable,[54] propensity score calibration,[55,56] pseudo treatment,[57] Rosenbaum sensitivity analysis,[58,59] Rosenbaum-Rubin sensitivity analysis,[60,61] and the trend-in-trend method.[62,63] However, these methods are more complicated in implementing, and require additional assumptions. Therefore, they may not be applicable for certain research scenarios. In addition, no single method would provide the best performance for each individual study. Thus, we discuss some factors that could help select appropriate methods in practice.

Previously we introduced the three possible scenarios concerning the availability of extra information on unmeasured confounders, which could be the first factor to consider when we plan the subsequent sensitivity analysis. The researchers should thoroughly think about the confounding factors and their availability in the databases and other sources (data, literatures, etc) before finalizing the analysis plan. Such information should drive the selection of available methods. Based on this factor, we categorized these analytical methods into three different categories, as shown at the last step of the flowchart.

The second factor is the type of unmeasured confounders. As discussed earlier, there are measurable confounders and unmeasurable confounders, and some analytical methods are only able to address measurable confounders but not unmeasurable ones. Multiple imputations, propensity score calibration, martingale residual-based method and Bayesian twin regression modeling are applicable when the unmeasured confounder is not available in the database butis measurable. Other methods could be implemented whether or not the unmeasured confounders is measurable.

The third factor is whether the subsequent sensitivity analysis is intended to provide an approximation of the causal treatment effect or rather to assess sensitivity to unmeasured confounding in a different manner. The former methods, which we term “direct adjustment methods”, approximate causal effect estimates generally by incorporating internal or external data or by invoking more statistical assumptions than used in the initial sensitivity analysis methods. Not exhaustively, some examples of direct adjustment methods are as follows. First, if internal data on the unmeasured

confounder(s) are available for some subjects, multiple imputation can be used to impute values of unmeasured confounders for the remaining subjects, and the causal treatment effect can be estimated using the imputed data. Second, propensity score calibration estimates the treatment effect using estimated propensity scores from the entire sample given the measured confounders only and also using estimated propensity scores from a random subset of subjects for whom the unmeasured confounders are in fact available. Third, Bayesian twin regression can potentially leverage information from internal or external data on unmeasured confounders to elicit prior information in order to adjust the posterior for unmeasured confounding.

Other methods, while not able to estimate an adjusted treatment effect estimate directly, could provide other evidence on the influence of unmeasured confounders. We term these "indirect methods". Examples of such methods are as follows. First, Manski's partial identification provides bounds on the treatment effect estimate such that any causal effect size inside the bounds cannot be ruled out. Second, negative control methods involve pre-specifying "negative control outcomes" that are thought to be unaffected by the treatment or, alternatively, "negative control exposures" interventions that are thought to have minimum impact on the outcome of interest. If, contrary to expectation, there is a non-negligible association between the treatment and negative control outcomes or between the negative control exposures and the outcome while adjusting for measured confounders, this would suggest the existence of unmeasured confounders. Third, if there are multiple control groups that differ systematically with respect to the outcome distribution after controlling for all measured confounders, this suggests the presence of residual unmeasured confounding.

The three factors we have discussed aim to provide general considerations for researchers when they design their studies and, ideally, pre-specify the analysis plan. Furthermore, each method listed at the last stage of the flowchart may have additional assumptions that needs to be satisfied for its validity. We have summarized some of these additional assumptions in Zhang et al[23], which we hope could further help exclude some methods when researchers pre-specify their analysis plans with regard to unmeasured confounding. Of course, there is probably no single method that always works better than other available ones as a subsequent sensitivity analysis. Thus, if multiple analytical methods are applicable given, for example, the nature of any substantive knowledge on unmeasured confounding and the availability of internal or external data, then it may be informative to apply more than one method (as subsequent sensitivity analyses) and report results from all conducted methods. Reporting of study conclusions should also then consider results from all those conducted sensitivity analyses.

So far, we have discussed the importance of controlling unmeasured confounder(s) in assessing causal inference from RWD. We also proposed a flowchart and some practical recommendations to conduct quantitative evaluation of the impact of unmeasured confounding on estimated treatment effect. The flowchart illustrates one possible structured approach, in the authors' opinions, to quantify and perhaps further adjust for the impact of unmeasured confounding. More importantly, it provides a tool to pre-specify the sensitivity analysis plan for unmeasured confounding when researchers are developing the protocol and statistical analysis plan for their own studies. Transparent, pre-specified and well-documented statistical conduct is essential to the credibility of scientific evidence because it ensures the methods can be analytically reproduced to confirm the findings, and it limits selective reporting and publication bias. These principles have been applied to randomized clinical trials for a long time, and are equally important for non-interventional studies in order to generate more credible and reliable RWE. Good procedural practices to enhance transparency and reproducibility for treatment effectiveness studies has been endorsed by relevant research societies,[64,65] and the FDA will consider those recommendations when issuing guidance about RWE from non-interventional studies to support product effectiveness in regulatory decision-making.[2] In our opinion, the proposed flowchart here also has the potential to contribute to such efforts to further improve the quality of RWE. Our approach will perhaps not be the only (or even the best) one to address this concern, but it certainly provides a practically useable tool regarding the analysis plan pre-specification. On the other end, while we feel our approach is reasonable and easy to implement in practice, we do acknowledge there is a great need for further systematic simulation-based analyzes of validation studies/real world studies and comparing those novel methods (and even comparing different "flowcharts") to generate more evidence for an improved systematic approach.

As a closing note, pharmacoepidemiologists, statisticians, and data scientists will continue to find ways to reduce confounding and other biases inherent in non-interventional RWD-based studies, including unmeasured confounding. Unmeasured confounding remains an important challenge for causal inference in non-randomized studies. Nevertheless, with improved statistical methods, more complete RWD (via linkage and novel data collection approach), and improved study designs (eg, self-controlled designs, positive/negative controls), our understanding regarding the impact of unmeasured confounding will continue to improve. Addressing unmeasured confounding to the fullest extent possible is critical to improving the quality of RWE for regulatory decisions by obtaining more reliable RWE from the explosion of RWD.

ACKNOWLEDGEMENTS

The authors are very grateful to Douglas E. Faries, four anonymous reviewers, and the associate editor for their valuable comments.

CONFLICT OF INTEREST

X. Z. was a full time employee of Eli Lilly and Company during the development of this manuscript. J. D. S. is a professor of statistics at Baylor University. M. B. M. is an assistant professor at the Quantitative Sciences Unit at Stanford University.

PRIOR PRESENTATIONS

Part of this work was presented at the 3rd International Conference on Pharmacoepidemiology & Therapeutic Risk Management and the 12th International Conference on Health Policy Statistics.

ORCID

Xiang Zhang https://orcid.org/0000-0002-5304-0482
James D. Stamey https://orcid.org/0000-0002-3787-6490
Maya B. Mathur https://orcid.org/0000-0001-6698-2607

REFERENCES

1. Attempt to replicate clinical trials with real-world data generates real-world criticism, too *STAT News*. July 3, 2019. Available from https://www.statnews.com/2019/07/03/replicate-clinical-trials-real-world-evidence/
2. How real world evidence was used to support approval of Ibrance for male breast cancer. *The Cancer Letter*. April 19, 2019. Available from https://cancerletter.com/articles/20190419_2/
3. European Medicines Agency. Guidance for companies considering the adaptive pathways approach. 2016. Available from https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guidance-companies-considering-adaptive-pathways-approach_en.pdf. Accessed December 1, 2019.
4. U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Dent Program 2018. https://www.fda.gov/media/120060/download. Accessed December 1, 2019.
5. Wedam S, Fashoyin-Aje L, Bloomquist E, et al. FDA approval summary: Palbociclib for male patients with metastatic breast cancer. *Clin Cancer Res*. 2019;26(5):1208-1212.
6. Grodstein F, Stampfer MJ, Manson JE, et al. Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. *New Engl J Med*. 1996;335(7):453-461.
7. 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.
8. Prentice RL, Langer RD, Stefanick ML, et al. Combined analysis of Women's health initiative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *Am J Epidemiol*. 2006;163(7):589-599.
9. Vandenbroucke JP. The HRT controversy: observational studies and RCTs fall in line. *Lancet*. 2009;373(9671):1233-1235.
10. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative—a comprehensive approach to medical product surveillance. *Clin Pharmacol Ther*. 2016;99(3):265-268.
11. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clin Pharmacol Ther*. 2019;105(4):867-877.
12. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945-960.
13. Causality PJ. *Models, Reasoning and Inference*. 2nd ed. New York, NY: Cambridge University Press; 2009.
14. Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. *JAMA*. 2016;316(17):1818-1819.
15. Shrank WH, Patrick AR, Brookhart MA. Healthy user and related biases in observational studies of preventive interventions: a primer for physicians. *J Gen Intern Med*. 2011;26(5):546-550.
16. Dormuth CR, Patrick AR, Shrank WH, et al. Statin adherence and risk of accidents: a cautionary tale. *Circulation*. 2009;119(15):2051-2057.
17. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf*. 2006;15(5):291-303.
18. 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.
19. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology*. 2003;14:451-458.
20. McMahon AD. Approaches to combat with confounding by indication in observational studies of intended drug effects. *Pharmacoepidemiol Drug Saf*. 2003;12(7):551-558.
21. Uddin MJ, Groenwold RH, Ali MS, et al. Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *Int J Clin Pharm*. 2016;38(3):714-723.
22. Streeter AJ, Lin NX, Crathorne L, et al. Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *J Clin Epidemiol*. 2017;87:23-34.
23. Zhang X, Faries DE, Li H, Stamey JD, Imbens GW. Addressing unmeasured confounding in comparative non-interventional research. *Pharmacoepidemiol Drug Saf*. 2018;27(4):373-382.
24. Smith GD, Ebrahim S. Data dredging, bias, or confounding. *BMJ*. 2002;325(7378):1437-1438.
25. ICH Topic E9 Statistical Principles for ClinicalTrials (CPMP/ICH/363/96). Available from www.emea.europa.eu. 1998. Accessed July 2, 2020.
26. Loder E, Groves T, Macauley D. Registration of observational studies. *BMJ*. 2010;340:c950.
27. The Lancet. Should protocols for observational research be registered? *Lancet*. 2010;375(9712):1.
28. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med*. 2017;167:268-274.
29. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91(434):444-455.
30. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. *Stat Med*. 2014;33(13):2297-2340.
31. Baiocchi M, Small DS, Lorch S, Rosenbaum PR. Building a stronger instrument in an observational study of perinatal care for premature infants. *J Am Stat Assoc*. 2010;105(492):1285-1296.
32. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf*. 2010;19(6):537-554.
33. MacKenzie TA, Tosteson TD, Morden NE, Stukel TA, O'Malley AJ. Using instrumental variables to estimate a Cox's proportional hazards regression subject to additive confounding. *Health Serv Outcomes Res Methodol*. 2014;14(1–2):54-68.
34. Cook TD. "Waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics. *J Econom*. 2008;142(2):636-654.
35. Imbens GW, Lemieux T. Regression discontinuity designs: a guide to practice. *J Econom*. 2008;142(2):615-635.
36. Mathur MB, Ding P, Riddell CA, VanderWeele TJ. Website and R package for computing E-values. *Epidimiology*. 2018;29(5):e45-e47.
37. VanderWeele TJ, Ding P, Mathur M. Technical considerations in the use of the E-value. *J Causal Inference*. 2017;7(2).1–11.
38. VanderWeele TJ, Ding P, Mathur MB. Developing best-practice guidelines for the reporting of E-values. *Int J Epidemiol*. 2020.
39. McCandless LC, Gustafson P, Levy A. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Stat Med*. 2007;26:2331-2347.
40. Zhang X, Faries DE, Boytsov N, Stamey JD, Seaman JWA Jr. Bayesian sensitivity analysis to evaluate the impact of unmeasured confounding with external data: a real world comparative effectiveness study in osteoporosis. *Pharmacoepidemiol Drug Saf*. 2016;25(9):982-992.
41. Abadie A. Semiparametric difference-in-differences estimators. *Rev EconomStud*. 2005;72(1):1-9.
42. Imbens GW, Wooldridge JM. Recent developments in the econometrics of program evaluation. *J Econom Lit*. 2009;47(1):5-86.
43. Norén GN, Bergvall T, Ryan PB, Juhlin K, Schuemie MJ, Madigan D. Empirical performance of the calibrated self-controlled cohort analysis within temporal pattern discovery: lessons for developing a risk identification and analysis system. *Drug Saf*. 2013;36(1):107-121.
44. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat Med*. 2014;33(2):209-218.

45. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidimiology*. 2009;20(4):512-522.
46. Manski CF. Nonparametric bounds on treatment effects. *Am Econ Rev*. 1990;80(2):319-323.
47. Burne RM, Abrahamowicz M. Martingale residual-based method to control for confounders measured only in a validation sample in time-to-event analysis. *Stat Med*. 2016;35(25):4588-4606.
48. Burne RM, Abrahamowicz M. Adjustment for time-dependent unmeasured confounders in marginal structural cox models using validation sample data. *Stat Methods Med Res*. 2019;28(2):357-371.
49. Abrahamowicz M, Bjerre LM, Beauchamp ME, LeLorier J, Burne R. The missing cause approach to unmeasured confounding in pharmacoepidemiology. *Stat Med*. 2016;35(7):1001-1016.
50. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf*. 2008;17(6):546-555.
51. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivar Behav Res*. 1998;33(4):545-571.
52. Lipsitch M, Tchetgen ET, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidimiology*. 2010;21(3):383-388.
53. Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? *JAMA*. 2013;309(3):241-242.
54. Lee WC. Detecting and correcting the bias of unmeasured factors using perturbation analysis: a data-mining approach. *BMC Med Res Methodol*. 2014;14(1):18.
55. Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol*. 2005;162(3):279-289.
56. Stürmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration—a simulation study. *Am J Epidemiol*. 2007;165(10):1110-1118.
57. Rosenbaum PR. The role of a second control group in an observational study. *Stat Sci*. 1987;2(3):292-306.
58. Gastwirth JL, Krieger AM, Rosenbaum PR. Asymptotic separability in sensitivity analysis. *J R Stat Soc Series B Stat Methodol*. 2000;62(3):545-555.
59. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*. 1987;74(1):13-26.
60. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J R Stat Soc B Methodol*. 1983;45(2):212-218.
61. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*. 1998;54:948-963.
62. Ji X, Small DS, Leonard CE, Hennessy S. The trend-in-trend research design for causal inference. *Epidimiology*. 2017;28(4):529-536.
63. Shahn Z. Trends in control of unobserved confounding. *Epidemiology*. 2017;28:537-539.
64. Berger ML, Sox H, Willke RJ, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE special task force on real-world evidence in health care decision making. *Pharmacoepidemiol Drug Saf*. 2017;26(9):1033-1039.
65. Wang SV, Schneeweiss S, Berger ML, et al. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1. 0. *Pharmacoepidemiol Drug Saf*. 2017;20(8):1009-1022.

How to cite this article: Zhang X, Stamey JD, Mathur MB. Assessing the impact of unmeasured confounders for credible and reliable real-world evidence. *Pharmacoepidemiol Drug Saf*. 2020;1–9. https://doi.org/10.1002/pds.5117