*Pharmaceutical Statistics*

WILEY

MAIN PAPER

# Real Effect or Bias? Good Practices for Evaluating the Robustness of Evidence From Comparative Observational Studies Through Quantitative Sensitivity Analysis for Unmeasured Confounding

Douglas Faries[1] | Chenyin Gao[2] | Xiang Zhang[3] | Chad Hazlett[4] | James Stamey[5] | Shu Yang[2] | Peng Ding[6] | Mingyang Shan[1] | Kristin Sheffield[7] | Nancy Dreyer[8]

[1]Real-World Access and Analytics, Eli Lilly & Company, Indianapolis, USA | [2]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA | [3]Medical Affairs Biostatistics, CSL Behring, King of Prussia, USA | [4]Departments of Statistics & Data Science and Political Science, University of California at Los Angeles, Los Angeles, USA | [5]Department of Statistical Science, Baylor University, Waco, USA | [6]Department of Statistics, University of California Berkeley, Berkeley, USA | [7]Value, Economics, and Outcomes, Eli Lilly & Company, Indianapolis, USA | [8]Dreyer Strategies, USA

**Correspondence:** Douglas Faries (fariesdouglas@gmail.com)

**Received:** 22 December 2023 | **Revised:** 24 August 2024 | **Accepted:** 14 November 2024

**Funding:** This work was supported by Eli Lilly and Company.

**Keywords:** bias | causal inference | confounding | real-world evidence | sensitivity

## ABSTRACT

The assumption of "no unmeasured confounders" is a critical but unverifiable assumption required for causal inference yet quantitative sensitivity analyses to assess robustness of real-world evidence remains under-utilized. The lack of use is likely in part due to complexity of implementation and often specific and restrictive data requirements for application of each method. With the advent of methods that are broadly applicable in that they do not require identification of a specific unmeasured confounder—along with publicly available code for implementation—roadblocks toward broader use of sensitivity analyses are decreasing. To spur greater application, here we offer a good practice guidance to address the potential for unmeasured confounding at both the design and analysis stages, including framing questions and an analytic toolbox for researchers. The questions at the design stage guide the researcher through steps evaluating the potential robustness of the design while encouraging gathering of additional data to reduce uncertainty due to potential confounding. At the analysis stage, the questions guide quantifying the robustness of the observed result and providing researchers with a clearer indication of the strength of their conclusions. We demonstrate the application of this guidance using simulated data based on an observational fibromyalgia study, applying multiple methods from our analytic toolbox for illustration purposes.

## 1 | Introduction

The growing availability of real-world data (RWD) has driven the use of real-world evidence (RWE) in the drug development and commercialization process, from discovery to phase IV research and market access. This has been spurred by the 21st Century Cures Act and subsequent efforts considering the use of RWE to inform regulatory decisions regarding the effectiveness and safety of medical products. However, the promise of timely and credible RWE to inform regulators and healthcare decision makers is challenged by the need to address potential biases inherent in non-randomized research. This is especially challenging for evidence derived from comparative observational studies—in which researchers often use causal inference methods to compare outcomes between 2 or more interventions [1]. This particular subset of RWE is the primary focus of this manuscript.

Generating credible estimates of causal treatment effects from observational studies requires making four key assumptions: the stable unit treatment value assumption (SUTVA), positivity, correct statistical modeling, and strong ignorability (or "no unmeasured confounders"). The "no unmeasured confounders" assumption is not verifiable and is a major roadblock to the acceptance of such evidence for healthcare decision making [2, 3]. An unmeasured confounder is a variable that is related to both the treatment and the outcome—yet is not available in the data set for analysis [4, 5]. Unmeasured confounding can be problematic, even reversing the direction and significance of treatment effect estimates [6]. Despite this, common practice is simply to discuss potential for bias due to unmeasured confounders as a limitation without quantitative sensitivity analyses [7, 8]. Best practice guidance for RWE, produced by the International Society of Pharmacoeconomics and Outcomes Research, the International Society of Pharmaoepidemiolog and others [7, 9–11], emphasize the importance of addressing unmeasured confounding but do not provide a roadmap for implementation. Similarly, study design good practices [12] do not address evaluation of robustness after study design.

Several books/review articles [13–16] provide summaries including the applicability and pros and cons of methods for addressing unmeasured confounding but are under-utilized for multiple reasons. First, many methods are complex and require custom programming for implementation. Second, many methods are only applicable in specific settings and not broadly applicable [15]. For instance, propensity score calibration requires a subsample of patients with data on the unmeasured confounder and the prior rate ratio requires outcome data in a time-period prior to when the exposure of interest started.

Recently, software to implement methods that are broadly applicable (requiring no knowledge of a specific unmeasured confounder or additional data) have become publicly available such as the R-packages TreatSens, Sensmakr, *E*-value, and Tipr. Researchers have proposed such methods as a first step for sensitivity analyses for comparative observational research [17–20]. Zhang and colleagues [18] developed a flowchart to help researchers navigate the options and produce sensitivity analyses appropriate for their study. However, this has not undergone assessment via pilot studies and only addresses sensitivity at the analysis stage (not the design stage). Also, expanded analytical options are available since that time.

Building on the work of Zhang et al. [18], we propose a good practices guidance for both the design and analysis stages of comparative observational research, including guiding questions and a toolbox of methods to help researchers quantitatively address the potential for unmeasured confounding. We illustrate the use of the guidance using simulated data based on a prospective real-world study of fibromyalgia [21] (REFLECTIONS).

## 2 | Sensitivity Analysis Guidance and Toolbox for Study Design and Analyses

### 2.1 | Toolbox Overview

Prior to introducing the sensitivity analysis guidance and toolbox in Sections 2.2 and 2.3, Table 1 provides background statistical details for methods in the toolbox and utilized in the example applications that follows. This includes three methods for tipping point analyses and benchmarking (*E*-value, Omitted Variables, Simulation Framework) along with approaches for adjusted analyses using internal or external data (Bayesian Regression, Control Variable Analysis). For notation across methods, Y represents the outcome, X the measured covariates, U the unmeasured covariates, and A the treatment.

### 2.2 | Study Design Stage

The need for planning sensitivity analyses for potential bias begins at the design stage of comparative observational research. Prior to developing a protocol, researchers should ensure the planned design and data will support a robust conclusion. Figure 1 provides a structured process to guide researchers to identify potential confounders, quantify the level of confounding that would be problematic to generating robust findings, conduct a benchmarking exercise based on the expectations, and consider options to reduce uncertainty caused by potential unmeasured confounding. To assist with implementation, a toolbox of analytic methods is included. Study objectives and designs vary, but these questions are broadly applicable, such as for non-inferiority, superiority, and predictive study aims, while the application of the toolbox may differ.

Girman et al. [32] recommended directed acyclic graphs (DAGs) to guide pre-study feasibility assessment in comparative real-world studies. Use of DAGs [33] at the design stage provides a structured approach to identify all known "common-cause confounders" (factors influencing both treatment selection and outcome)—measured or not—based on current evidence. Information to develop DAGs is obtained from sources such as clinical experts, prescriber surveys, literature reviews, and existing disease state data. Tools like DAGitty [34] can be utilized to develop or analyze DAGs. See Digitale et al. [35] for a recent tutorial while Ferguson et al. [36] discuss recent approaches and challenges for building DAGs from prior information and Khuene et al. [37] provides an applied example in a causal inference setting. The simple DAG in Figure 2 for a single treatment decision point illustrates that bias from unmeasured confounding is driven by two factors (after conditioning on X): (1) the strength of association between the unmeasured confounder U and treatment A; (2) the strength of association between the unmeasured confounder U and outcome Y.

Once potential unmeasured confounders are identified, researchers need to make informed assumptions of the magnitude of the correlation between the unmeasured confounders and outcome and/or treatment (adjusted for other factors) and the expected treatment effect. The *E*-value and robustness

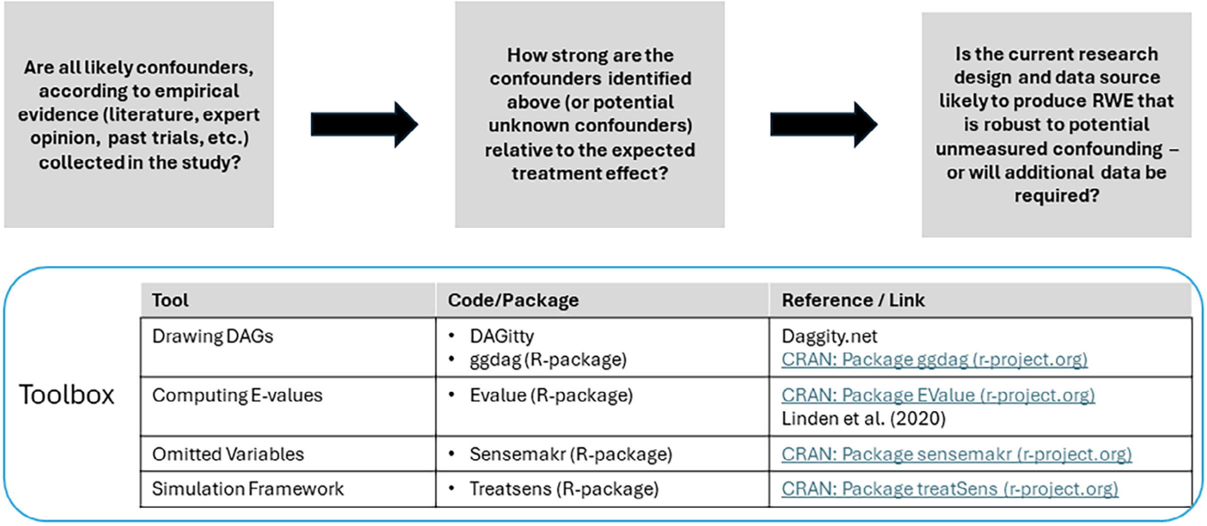
**TABLE 1** | Methods overview.

| Method | Description |
|---|---|
| *E*-value | The *E*-value, representing the evidence for causality, is the minimum strength of confounding necessary on a risk ratio (RR) scale that an unmeasured confounder would need to have with both the treatment and outcome to explain away a specific observed treatment effect, conditional on the observed covariates [17].<br>For the technical derivation see VanderWeele, Ding, and Mather [22] where they show how the *E*-value is based on two parameters:<br>$RR_{UY} = max\left( \frac{max_u P(Y=1 \mid A=1, X=x, U=u)}{min_u P(Y=1 \mid A=1, X=x, U=u)}, \frac{max_u P(Y=1 \mid A=0, X=x, U=u)}{min_u P(Y=1 \mid A=0, X=x, U=u)} \right)$ and<br>$RR_{AU} = max_u \frac{P(U=u, A=1, X=x)}{P(U=u \mid A=0, X=x)}$<br>$RR_{UY}$ is the maximum effect that U can have on Y, conditional on X = x and $RR_{AU}$ is the maximum RR for the treatment A across possible levels of U.<br>For a binary outcome on a RR scale (also with a binary unmeasured confounder), the *E*-value for the point estimate can be computed as follows:<br>If RR ≥ 1: $\text{Evalue} = RR + \sqrt{RR(RR-1)}$<br>If RR < 1: $Evalue = \frac{1}{RR} + \sqrt{\frac{1}{RR}\left(\frac{1}{RR} - 1\right)}$<br>The *E*-value for the confidence limit of interest (LL = Lower Limit; UL = Upper Limit) would be<br>If RR ≥ 1: $\text{Evalue} = \begin{cases} LL + \sqrt{LL(LL-1)}, if\, LL > 1 \\ 1, \quad if\, LL \leq 1. \end{cases}$<br>If RR < 1: $\text{Evalue} = \begin{cases} \frac{1}{UL} + \sqrt{\frac{1}{UL}\left(\frac{1}{UL} - 1\right)}, if\, UL < 1 \\ 1, \quad if\, UL \geq 1. \end{cases}$<br>It is recommended to report the *E*-value for both the point estimate and the confidence interval limit closest to the null—thus creating a measure of the strength of confounding needed to change any inferences from the study.<br>Extensions of the *E*-value for continuous, time to event, and odds ratio outcomes are available and implementation of the *E*-value can be conducted using multiple packages including the *E*-value R-package [23, 24]. |
| Omitted variables | Cinelli and Hazlett proposed using the omitted variables framework with a partial $R^2$ parameterization to develop a suite of tools to assess the impact of unmeasured confounding [19]. Beginning with an omitted variables framework, the desired regression model is assumed to be<br>$Y = \beta X + \zeta_Y U + \tau A + \epsilon$<br>with outcome Y, covariates X (measured) and U (unmeasured), and treatment A. The analysis model, however, excludes the parameter U and Cinelli and Hazlett then compute the bias from the exclusion of U and parameterize the bias in terms of the proportion of variance explained ($R^2$).<br>$\lvert bias \rvert = \sqrt{\frac{R^2_{Y \sim U \mid A,X} \, R^2_{A \sim U \mid X}}{1 - R^2_{A \sim U \mid X}}} \left( \frac{sd(Y^{\neg X,A})}{sd(A^{\neg X})} \right),$<br>where $Y^{\neg X,A}$ is the variable Y after removing components linearly explained by X and A.<br>Consider a confounder U with equal association to the treatment and the outcome, then the Robustness Value, the minimum strength of association U would need with both the treatment and outcome to eliminate or change statistical inferences (reduce the estimated effect by 100 × q%), is<br>$RV_q = (0.5)\left(\sqrt{f_q^4 + 4f_q^2} - f_q^2\right)$<br>where $f_q$ is the partial Cohen's f of the treatment with the outcome multiplied by the proportion of reduction q of the treatment coefficient deemed of interest.<br>Confounders that explain this proportion of the residual variance for both the treatment and the outcome are sufficiently strong to change the point estimate in problematic ways, while confounders with neither association greater than this proportion are not.<br>They propose using sensitivity contour plots with the partial $R^2$ parameterization for benchmarking. Specifically, create a contour plot (see Section 3.2.2) of adjusted treatment effect estimates where the x-axis is the partial $R^2$ of the unmeasured confounder with treatment and the y-axis is the partial $R^2$ of the confounder with outcome. One can then use the partial $R^2$ values from the strongest measured confounder as a benchmark and assess on the contour plot whether an unmeasured confounder of similar strength (or 2-times, 3-times, ...) would make important changes to the point estimate and inferences from the study. |

(Continues)

**TABLE 1** | (Continued)

| Method | Description |
|---|---|
| Simulation framework | Under the simulation-based framework for sensitivity analysis [25, 26], parametric models are posited for the treatment assignment conditional on observed confounders and outcome given treatment assignment and observed confounders. For instance, a linear regression model can be specified for a continuous outcome and the treatment assignment can be modeled under a logistic regression. Sensitivity parameters $\zeta_Y$ and $\zeta_A$ are introduced to denote the association between the outcome and an unmeasured confounder and the treatment assignment and an unmeasured confounder, respectively. Finally, the unmeasured confounder is assumed to be binary with marginal proportion $\pi_U$, with $\pi_U = 0.5$ in the simulated example of Section 3. Let<br>$Y \mid X, A, U \sim N\left(\beta X + \zeta_Y U + \tau A, \sigma^2\right),$<br>$A \mid X, U \sim Bernoulli\left(expit\left(\psi X + \zeta_A U\right)\right)$, and<br>$U \sim Bernoulli\left(\pi_U\right).$<br>To conduct a tipping point analysis, a grid of values are specified for $\zeta_Y$ and $\zeta_A$. For each combination of $\left(\zeta_Y, \zeta_A\right)$, an algorithm iterates between (1) re-estimating the parameters for the outcome $\left(\beta, \tau, \sigma^2\right)$ given $Y, X, A, U, \zeta_Y$, (2) re-estimating the parameter for the treatment assignment $(\psi)$ given $A, X, U, \zeta_A$, and (3) the conditional probability of $\Pr\left(U = 1 \mid Y, A, X, \beta, \tau, \psi, \zeta_Y, \zeta_A\right)$ to obtain an adjusted estimate of the treatment effect. The simulation-based sensitivity analysis framework can be implemented in R via the package treatsens. Contour plots with benchmarking are also recommended to evaluate the robustness of inferences from the study. The x and y-axis on the countour plot reflect different values of $\zeta_Y$ and $\zeta_A$, which are on the scale of regression coefficients in the outcome and treatment models, respectively. |
| Bayesian Regression | The Bayesian approach is similar to missing data approaches where information for the unmeasured confounder is used to develop a posterior distribution for the missing variables [6, 27, 28]. Treating the values of the unmeasured confounders as unknown parameters then allows them to be imputed into the model as a part of the MCMC algorithm. Estimation proceeds on the full data set (with the imputed values) where uncertainty is injected into the parameter estimates due to the missing data being a part of the iterative sampling scheme. Parametric models for both the unmeasured confounder and the response are typically used,<br>$Y \mid A, X, U \sim D_Y\left(\theta_Y\right),$<br>$U \mid A, X \sim D_U\left(\theta_U\right).$<br>The means for U and Y are, for appropriate link functions, given as<br>$F(U \mid A, X) = \varphi_{U0} + \varphi_U^A X$<br>$G(Y \mid A, X, U) = \varphi_{Y0} + \varphi_Y^A X + \zeta_Y U.$<br>If internal or external validation data are available, inference can proceed with non-informative priors for all parameters. In the absence of validation data, informative priors are required for $\varphi_{U0}$, $\varphi_U^A$, and $\zeta_Y$ . The informative priors can be elicited either via independent normal priors for each of the parameters or a joint elicitation process using the conditional means prior approach [29]. The posterior distributions would rarely be of closed form, thus MCMC methods are typically used to perform estimation. The R-package unmconf is a user-friendly approach that can be used to fit these models via the JAGS software [30]. |
| Control variable | Yang and Ding [31] developed a methodology that integrates a main observational data set with unmeasured confounders ($U$) and a smaller internal validation data set that provides additional information on these confounders. The Control Variable approach first computes an initial internal estimator $\widehat{ATE}_0$ that adjusts for the known confounders and $U$, ensuring consistency but not guaranteed efficiency. The choices of the estimators can vary, e.g., inverse probability weighting, ANCOVA adjustment, augmented inverse probability weighting, and matching estimators. By adjusting for known confounders and $U$, $\widehat{ATE}_0$ is consistent for the average treatment effect. However, it might not be efficient due to using the smaller internal validation data set.<br>Secondly, the Control Variable approach determines an error-prone estimator $\widehat{ATE}^{ep}$ (where the superscript *ep* means "error-prone") utilizing both the main and validation data sets adjusting for known confounders but not $U$. When the validation data set is a simple random sample of the main data set, the difference between the error-prone estimators $\widehat{ATE}^{ep}_{val} - \widehat{ATE}^{ep}_{main}$ derived from both studies is consistent for zero. If this difference shares a strong correlation with the initial internal estimator for the ATE, $\widehat{ATE}_0$, it can be utilized as a control variate to improve the efficiency of the initial estimator. Thus, we call $\widehat{ATE}^{ep}_{val} - \widehat{ATE}^{ep}_{main}$ the control variate. The final estimator, by incorporating the control variate<br>$\widehat{ATE} = \widehat{ATE}_0 - \widehat{\Gamma}_{0c}\widehat{V}_c^{-1}\left(\widehat{ATE}^{ep}_{val} - \widehat{ATE}^{ep}_{main}\right),$<br>is both consistent and more efficient than the initial estimator, where $\widehat{\Gamma}_{0c}$ is the estimated correlation between the initial estimator $\widehat{ATE}_0$ and the control variate $\widehat{ATE}^{ep}_{val} - \widehat{ATE}^{ep}_{main}$, and $\widehat{V}_c$ is the estimated variance of the control variate. $\widehat{ATE}$ is guaranteed to be no worse than $\widehat{ATE}_0$ by borrowing the information from the control variate. |



Are all likely confounders, according to empirical evidence (literature, expert opinion, past trials, etc.) collected in the study?

How strong are the confounders identified above (or potential unknown confounders) relative to the expected treatment effect?

Is the current research design and data source likely to produce RWE that is robust to potential unmeasured confounding – or will additional data be required?

Toolbox

| Tool | Code/Package | Reference / Link |
|---|---|---|
| Drawing DAGs | • DAGitty<br>• ggdag (R-package) | Daggity.net<br>CRAN: Package ggdag (r-project.org) |
| Computing E-values | • Evalue (R-package) | CRAN: Package EValue (r-project.org)<br>Linden et al. (2020) |
| Omitted Variables | • Sensemakr (R-package) | CRAN: Package sensemakr (r-project.org) |
| Simulation Framework | • Treatsens (R-package) | CRAN: Package treatSens (r-project.org) |

**FIGURE 1** | Pre-study planning: Guiding questions and toolbox.

value (RV) are statistics that quantify what level of confounding could produce the observed treatment effect when the true effect is zero. While the treatment effect is not known prior to the study, researchers can contemplate an expected treatment effect size, which can be used to compute a pre-study *E*-value or RV (for both the expected treatment effect and confidence limit of interest).

Once bounds are computed and potential unmeasured confounders identified, benchmarking exercises compare the expected strength of confounders (based on clinical expertise or existing data) to the identified bounds of concern. Even if no specific unmeasured confounder has been identified, there is still potential bias from unknown confounders. In such cases a measured confounder serves as a proxy in the benchmarking process. That is, researchers may hypothesize that unknown confounders are not stronger (by a specified multiple) than proven known confounders and can use the strongest known confounder as a conservative benchmark.

This exercise may raise or lessen concerns with unmeasured confounding. When unmeasured confounding could have a significant impact on analyses then a discussion of options prior to moving forward with the study is warranted. While choices are situation dependent due to many factors, options include not conducting the study or at least the comparison piece of the study or obtaining additional information on the unmeasured confounder through alternative data collection approaches such as surveys, chart reviews or external data sources. Assuming the study continues, pre-specification of approaches to obtain additional information to reduce expected uncertainty from unmeasured confounding and how this information will be used in the analysis should be documented. In general, increasing the sample size of the study will not address the amount of bias in the treatment effect estimate. However, Fang et al. [38] note that while the expected effect size is fixed, the confidence limits vary with sample size. Thus, if the *E*-value or RV suggests that the study will lead to a point estimate that is arguably robust to plausible degrees of confounding, but the relevant confidence limit is not, then sample size considerations are relevant.

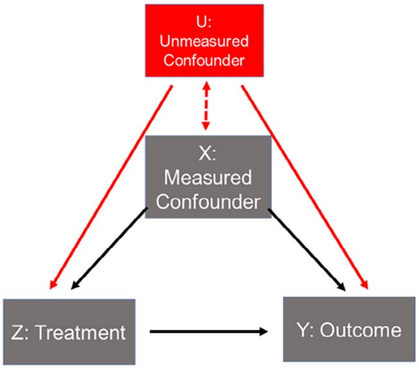


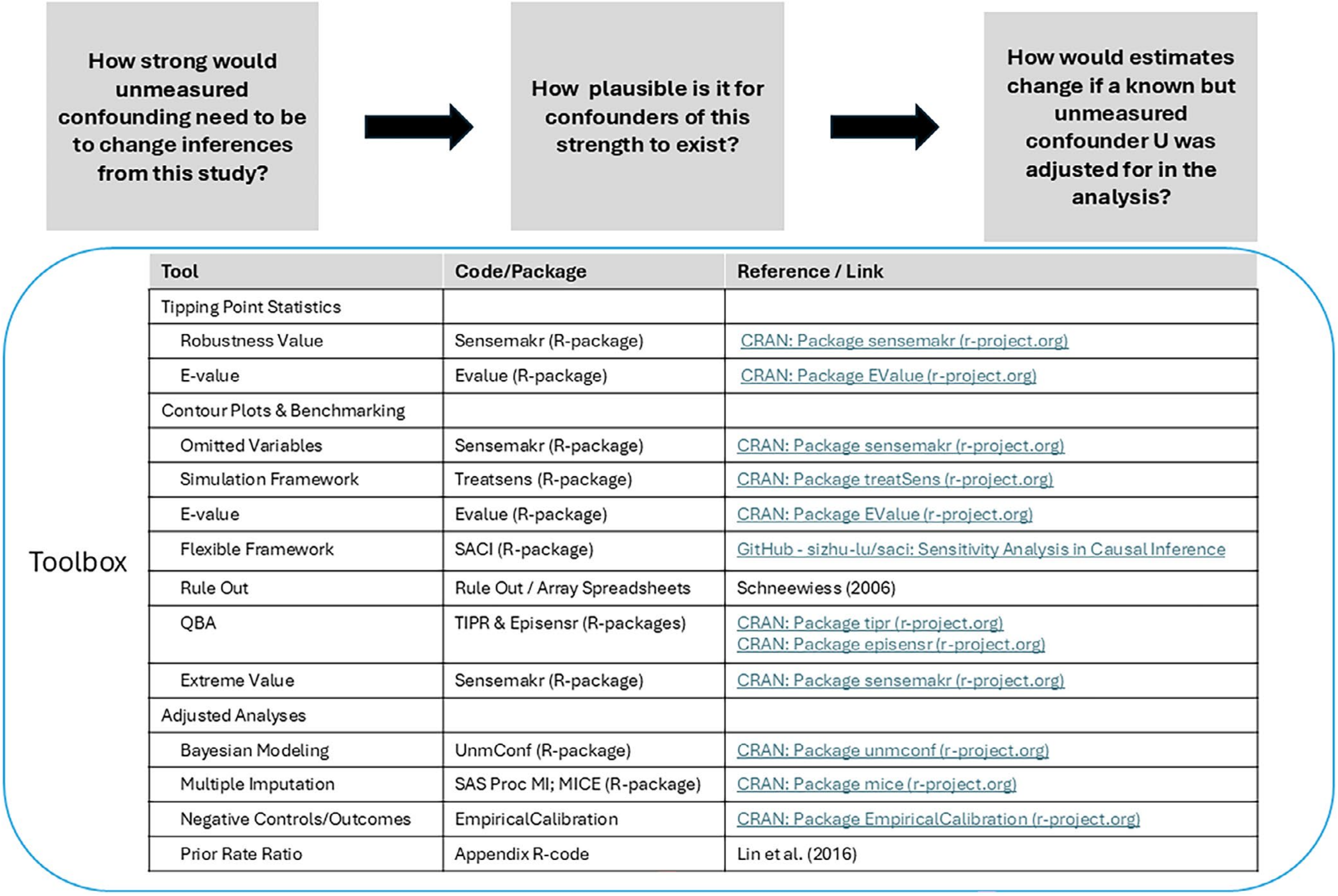
**FIGURE 2** | DAG summarizing measured and unmeasured confounding.

### 2.3 | Analysis Stage

After primary data analyses are completed, sensitivity analyses address how strong unmeasured confounding would need to be to change inferences. Figure 3 provides guidance and tools for quantitative sensitivity analyses to understand the robustness of the findings to potential unmeasured confounding.

As a general guidance, addressing the 1st two questions with tipping point and benchmarking analyses would be a minimum sensitivity assessment in most research scenarios. To address "How strong would unmeasured confounding need to be to change inferences from the study?", there are multiple tipping point analyses available including the *E*-value [17], omitted variables (including the RV and extremeRV) [19], simulation framework [25, 26] and array approach [39]. These methods are broadly applicable, given they require no knowledge of any specific confounder and approximations allow application for continuous, binary, and time to event outcomes. See Section 2.1 for some technical background information on these methods.

In addition to a single summary statistic such as the *E*-value or RV, we recommend contour plots for summarizing information



How strong would unmeasured confounding need to be to change inferences from this study?

How plausible is it for confounders of this strength to exist?

How would estimates change if a known but unmeasured confounder U was adjusted for in the analysis?

Toolbox

| Tool | Code/Package | Reference / Link |
|---|---|---|
| Tipping Point Statistics | | |
| Robustness Value | Sensemakr (R-package) | CRAN: Package sensemakr (r-project.org) |
| E-value | Evalue (R-package) | CRAN: Package EValue (r-project.org) |
| Contour Plots & Benchmarking | | |
| Omitted Variables | Sensemakr (R-package) | CRAN: Package sensemakr (r-project.org) |
| Simulation Framework | Treatsens (R-package) | CRAN: Package treatSens (r-project.org) |
| E-value | Evalue (R-package) | CRAN: Package EValue (r-project.org) |
| Flexible Framework | SACI (R-package) | GitHub - sizhu-lu/saci: Sensitivity Analysis in Causal Inference |
| Rule Out | Rule Out / Array Spreadsheets | Schneeweiss (2006) |
| QBA | TIPR & Episensr (R-packages) | CRAN: Package tipr (r-project.org)<br>CRAN: Package episensr (r-project.org) |
| Extreme Value | Sensemakr (R-package) | CRAN: Package sensemakr (r-project.org) |
| Adjusted Analyses | | |
| Bayesian Modeling | UnmConf (R-package) | CRAN: Package unmconf (r-project.org) |
| Multiple Imputation | SAS Proc MI; MICE (R-package) | CRAN: Package mice (r-project.org) |
| Negative Controls/Outcomes | EmpiricalCalibration | CRAN: Package EmpiricalCalibration (r-project.org) |
| Prior Rate Ratio | Appendix R-code | Lin et al. (2016) |

**FIGURE 3** | Analysis stage guiding questions and toolbox.

from the tipping point analyses. Contour plots allow postulating different strengths of association (between the unmeasured confounder and treatment selection and between the unmeasured confounder and outcome) to examine the resulting change in the adjusted treatment effect or any change in statistical inferences. They avoid the simplification of assuming an unmeasured confounder with the same level of influence on both the treatment and outcome and serve as a useful tool in the benchmarking analyses that follow. The different methods utilize different scales for measuring the associations with the unmeasured confounder (RR for the $E$-value, $R^2$ for the omitted variables, and regression coefficient for the simulation framework), but the concept is similar.

Once bounds are established for the strength of confounding that would change inferences, the plausibility of the existence of confounders of this strength should be considered. Benchmarking can be helpful to address the question. In benchmarking, researchers postulate various plausible strengths of associations (between the unmeasured confounder and both outcome and treatment selection) based on expert knowledge or external information (prior studies or other data sets) on a known unmeasured confounder or using data from measured covariates in the study when no unmeasured confounder has been identified. Contour plots can then be used to transparently visualize how treatment effect estimates or statistical inferences would change subject to any given benchmark. If the user finds that plausible benchmark values would not alter the research conclusion, this can be very informative. Alternatively, if researchers find confounding as strong as plausible benchmarks would overturn the research conclusions, then they know that confidence in the research conclusion is not warranted.

In some cases, sensitivity analysis only will reveal that we cannot be certain to even the sign of the true effect. In such cases, it may be useful to seek out additional information from other data sources that would aid in benchmarking exercises. Alternatively, entirely different causal identification strategies such as instrumental variables [40] and exploiting information in negative controls/outcomes [41, 42] may be more productive, depending on whether their alternative assumptions are defensible.

In cases where an unmeasured confounder U is identified and additional information regarding U is available, several approaches are available to assess whether inferences would change if U could be adjusted for in the analysis. Lash et al. [16] introduced a broad framework, quantitative bias analysis, which incorporate external information to estimate the potential impact of unmeasured confounders, as well as sensitivity analyses for missing responses, selection bias, exposure, outcome and covariate misclassification. The set of tools for incorporating additional information into sensitivity analyses is growing, though they are dependent on the type of information that is available for the unmeasured confounder U. Several recent approaches for both internal and external information are described below.

If information on the unmeasured confounder U is available from an internal subsample (a subset of the patients in the current study), then missing data methods such as multiple imputation and the Control Variable approach [31] (see Section 2.1) become feasible. Multiple imputation generates imputations of U based on the joint distribution of all confounders, treatment, and outcome under the missing at random assumption. For each imputation the treatment effect is re-estimated (using the imputed value of U) and estimates are pooled for a final estimate [43]. The Control Variable approach first computes an initial internal estimator that adjusts for the observed confounders and U within the subgroup where U is observed, which is consistent but not efficient. To improve efficiency a weighted estimator is obtained by combining information from the subset and the potential error-prone full data set.

When information on an unmeasured confounder is available from data external to the current study, Bayesian methods are a useful tool as they are designed to incorporate multiple sources of information [27, 28]. As described in more detail in Section 2.1, the external information—such as measures of the relationship between U and treatment and/or outcome—can be incorporated into the treatment effect estimation through prior distributions for the parameters governing the unmeasured confounder. Zhang et al. [6] provide an example of using Bayesian modeling to incorporate external information on a missing confounder into a claims-database analysis and demonstrated that analyses without the known but unmeasured confounders was likely significantly biased.

Lastly, the contour plots produced by the omitted variables or simulation framework can be paired with any additional information on U (whether internal or external) to provide an updated treatment effect estimate. One simply needs to obtain or postulate the strengths of association of U with both treatment and outcome from the additional data source. Note that these methods utilize additional information on U to provide updated treatment effect estimates. When applied to a single “U” of concern, this does not address other potential unmeasured confounders. However, contour plots (and the RV or *E*-value) can also be used to speculate about the strength of all confounding collectively. At any speculated values of the residual variance in the treatment and in the outcome that could be explained by all confounders collectively, the bias formulas describe the maximum bias due to such confounding.

## 3 | Pilot Application Using Simulated Study Data on Fibromyalgia

We piloted the good practice guidance and toolbox using our simulated REFLECTIONS fibromyalgia study. REFLECTIONS was a prospective observational study [21] that enrolled 1700 patients initiating a new treatment for fibromyalgia (new user design) and collected data longitudinally to compare 1-year pain severity outcomes (Brief Pain Inventory [BPI], null hypothesis of no treatment difference) in patients initiating opioid versus non-opioid treatments [44]. To have an example with known levels of unmeasured confounding and known true treatment effect, we generated a simulated version of the REFLECTIONS data ($N = 1000$) with:

- The same set of baseline covariates with the same distributions and correlations as in the actual study;
- A newly created variable “U” which represents an unmeasured confounder;
- New treatment (Opioid or Non-Opioid) and outcome (pain severity) variables such that there was no true treatment effect and U was related to both treatment selection and outcome with strengths similar to baseline BPI.

Details on the data generation process are presented in the Appendix S1. To demonstrate the use of sensitivity analysis methods that leverage external data, we also generated a fibromyalgia disease registry data set with similar covariates and outcomes as in REFLECTIONS—but only with non-opioid treated patients.

### 3.1 | Design Stage

#### 3.1.1 | Identify Measured and Unmeasured Confounders

Study design followed the analysis framework of Ho et al. [45] with details in the Appendix S1. Following the principles in Figure 1, we first created a DAG to assess whether all known confounders were collected in the study (Figure 4). For simplicity, we focused on the relationships of each covariate with treatment selection and 1-year pain severity. While no known unmeasured confounders were identified, potential for unmeasured confounding bias still exists as the lack of known unmeasured confounders is not proof of no unmeasured confounding.

#### 3.1.2 | Assess Strength of Unmeasured Confounding

Prior evidence and clinical insight suggest that baseline pain severity was likely the strongest confounder, meaning it is unlikely that unmeasured confounding from unknown sources would be stronger than confounding from baseline pain severity. Thus, baseline pain severity served as a conservative benchmark for potential unknown confounding in the sensitivity analysis below.

To assess the robustness of the study design against unknown confounding benchmarked at the level of “baseline pain severity,” a pre-study *E*-value was computed. Based on the 0.25 expected treatment effect, a standard deviation of 1.2, and planned sample sizes of 200/400 per group, the *E*-values for the point estimate and lower confidence limit would be 1.82 and 1.31. The external registry RR for baseline pain severity with pain severity outcome was 3.1. Given this was greater than 1.31, the study design was considered not likely to produce RWE robust against unmeasured confounding at the level of strength of baseline pain severity.

#### 3.1.3 | Consider Additional Sources of Information for the Unmeasured Confounder

Given the benchmarking results, we proceeded to the final pre-study step of considering opportunities to collect additional

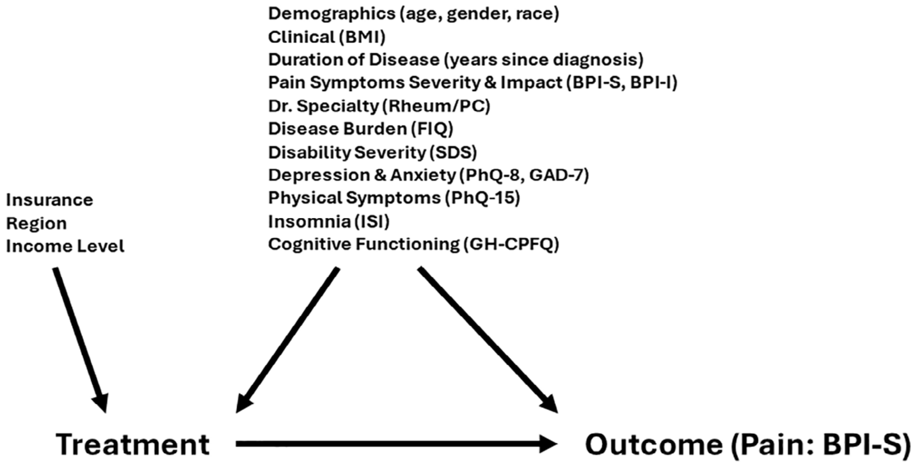

Demographics (age, gender, race)
Clinical (BMI)
Duration of Disease (years since diagnosis)
Pain Symptoms Severity & Impact (BPI-S, BPI-I)
Dr. Specialty (Rheum/PC)
Disease Burden (FIQ)
Disability Severity (SDS)
Depression & Anxiety (PhQ-8, GAD-7)
Physical Symptoms (PhQ-15)
Insomnia (ISI)
Cognitive Functioning (GH-CPFQ)
Insurance
Region
Income Level
Treatment
Outcome (Pain: BPI-S)

**FIGURE 4** | REFLECTIONS study DAG.

information to produce a more robust treatment effect estimate. For demonstration purposes, we assume a specific unmeasured confounder, denoted by U, was identified. Plans were then made to obtain patient level data on the unmeasured confounder at a limited number of investigational sites (e.g., chart reviews). These partial data are used below to provide additional sensitivity analysis.

## 3.2 | Analysis Stage

### 3.2.1 | Treatment Effectiveness Analysis

As detailed in the Appendix S1, we conducted Targeted Maximum Likelihood Estimation (TMLE) to estimate the effect of treatment on 1-year pain severity using only the measured baseline covariates to control for bias (ignoring the unmeasured confounder U). The estimand of interest was the average treatment difference in change in pain severity for the full population (ATE) assuming no further effect following early discontinuation of the treatment. Analyses were conducted using the TMLE R-package [46] and R-code is provided in the Appendix S1. Results showed a statistically significantly greater mean pain reduction in Opioid treated patients on the BPI-Pain scale (−0.37 [−0.64, −0.10]).

### 3.2.2 | Sensitivity Analysis: Tipping Point and Benchmarking Analyses

To assess the robustness of the findings, we began by evaluating how strong unmeasured confounding must be to change inferences from the study and assessing the plausibility of the existence of such confounding. For demonstration purposes we present 3 different tipping point and benchmarking analyses: the *E*-value, omitted variables, and simulation framework (R-code provided in the Appendix S1). For benchmarking we use our strongest measured confounder, baseline pain severity, as a priori we assumed an unmeasured confounder would not likely exceed the level of confounding produced by this variable.

Figure 5a–c displays the results. The *E*-values for the point estimate and lower confidence limit were 1.7 and 1.2. Thus, the existence of a binary unmeasured confounder with a RR of at least 1.2 with both treatment selection and pain severity could result in the statistically significant finding becoming non-significant. The benchmarking exercise shows the observed statistically significant result is not robust to unmeasured confounding at the strength of baseline pain severity (with a RR of 1.47 with treatment and 3.08 with outcome). That is, if an unmeasured confounder with the strength of confounding as baseline pain severity existed, the observed effect would no longer be statistically significant.

The omitted variables and simulation framework contour plots (Figure 5b,c) display combinations of strength of confounding that could fully explain the observed result (areas above and right of the red dashed line) or eliminate the statistical significance (areas above and right of the blue dashed line). Using the Sensmakr R-package, the Robustness values for the treatment effect and lower confidence limits were 0.077 and 0.012. Thus, an unmeasured confounder explaining only 1.2% of the residual variance in both the outcome and treatment models would imply that the treatment effect, adjusted for such confounding, would be consistent with the null hypothesis of no effect. As with the *E*-value approach, these benchmarking analyses show that, if unobserved confounding is as strongly related to treatment and outcome as baseline pain severity, then it would cut the point estimate nearly in half (from −0.30 to −0.17), at which point the estimate would not be statistically distinguishable from zero at the 95% confidence level.

### 3.2.3 | Sensitivity Analysis: Using Additional Information

Given the potential lack of robustness from the initial sensitivity analyses, we next consider whether incorporating additional information on the presumed unmeasured confounder, U, can provide greater clarity. As mentioned previously, the available analysis options depend upon the type of information about U that is available. Here we demonstrate analyses given (1) information on U in a separate study (external); (2) information on U from a subsample of the patients in the study (internal).

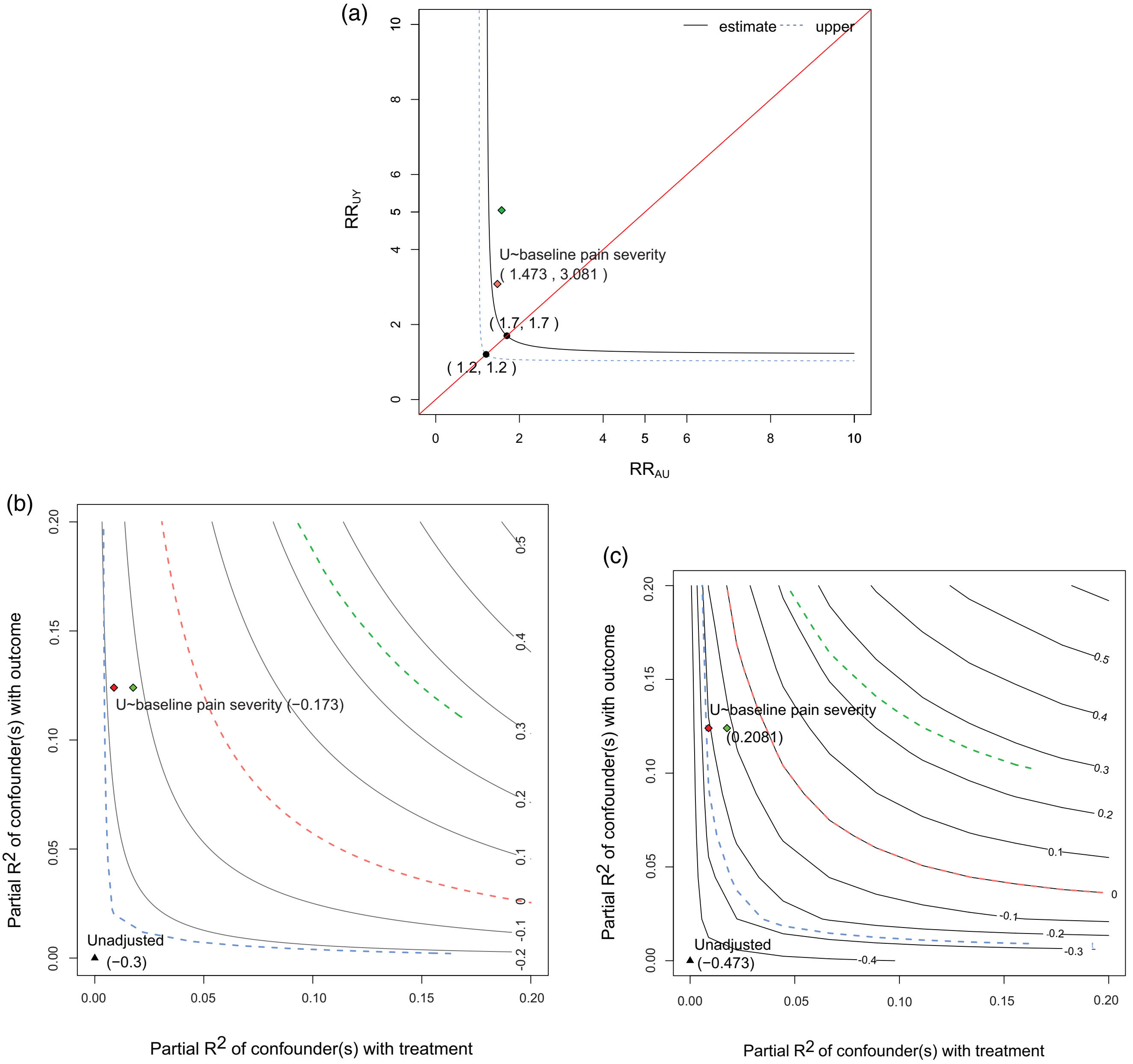


**FIGURE 5** | Tipping point sensitivity analysis boundary plots: (a) *E*-value, (b) omitted variables, and (c) simulation framework.

With information from U from a similar population of patients in the separate disease registry (see Appendix S1), contour plots, quantitative bias analyses [16, 47] and Bayesian approaches can be used to sharpen our sensitivity analyses. Suppose the external registry shows a strength of association between U and treatment of $R^2 = 0.022$ and between U and outcome of $R^2 = 0.122$. Plotting this external information onto the contour plots in Figure 5b,c suggests a true effect of approximately −0.10. Thus, the internal information successfully generated estimates closer to the null truth than either the original biased analyses or using the conservative "baseline severity" variable for benchmarking.

Lastly, we demonstrate the use of "internal" information on the unmeasured confounder for the sensitivity analysis. Table 2 displays the results of incorporating patient level information on U (here a subsample of data from select study sites amounting to 22% of the total sample) through multiple imputation, the Control Variable approach, and Bayesian regression modeling. Multiple imputation was performed using the MICE R-package and models leveraged all baseline and outcome information. For the Bayesian approach, the twin regression model was applied with relatively non-informative priors on all parameters. All three methods produced similar treatment effect estimates, ranging from −0.12 to −0.17. While the true data generating model had no treatment effect, the treatment effect estimate for our single random sample, when using the correct regression model with U and all other covariates in the model was −0.08. Thus, these internal methods produced estimates close to the best possible analysis from the data set, though not as close as the use of the contour plots and external information in the prior paragraph (−0.10).

**TABLE 2** | Summary of sensitivity analyses using internal information on U.

| Method | Adjusted treatment effect estimate | Standard error | 95% confidence interval |
|---|---|---|---|
| Multiple imputation | −0.15 | 0.14 | (−0.43, 0.13) |
| Control variable approach | −0.14 | 0.26 | (−0.65, 0.38) |
| Bayesian modeling | −0.17 | 0.12 | (−0.41, 0.07) |

### 3.2.4 | Summary of Pilot Study

In summary, the original analyses ignoring the unmeasured confounder suggested a statistically significantly greater reduction in pain severity for patients in the Opioid treatment group relative to the Non-Opioid group. However, following the sensitivity analyses suggested by our proposed guidance, we found the statistically significant result was not robust against unmeasured confounding at the strength of baseline pain severity. This was indicated by the low E- and Robustness values for the lower confidence limit, the benchmarking exercises with baseline pain severity, and with adjusted analysis incorporating internal and external information showing a smaller and non-significant treatment effect estimate. Clinical judgment is still required to consider the possibility of an unmeasured confounders of this strength in this situation. Regardless, incorporating additional information on a potential unmeasured confounder from either external or internal data sources produced much smaller treatment effect estimates close to the true value.

### 3.2.5 | Sample Size Simulation Study

A thorough comparison of the operating characteristics between various sensitivity analysis methods is complicated by the fact that different information is used by different methods and is beyond the scope of this work. However, important practical questions about the operating characteristics of this proposed guidance document remain. One critical step in all study planning is ensuring sufficient sample size to achieve study objectives. Given the importance of this issue, we developed a simulation study based on our pilot study to understand the impact of sample size on the operating characteristics of the proposed sensitivity guidance.

Data generation models and parameters for the simulation study were the same as for the pilot study, with details provided in the Appendix S1. Here we simulated 500 data sets with half, the same, and double the sample size as the single pilot example. For each simulated data set under the same data generation mechanism we then estimated the treatment effect ignoring the unmeasured confounding, computed an E- and Robustness value, performed benchmarking analysis, and computed an adjusted treatment effect using a 20% internal data sample using a control variable approach.

Table 3 presents the summary of the 500 simulations for the primary treatment comparison analyzed using TMLE. As expected, the treatment effect point estimate is largely unchanged as sample size changes. However, doubling the sample size produces narrower confidence intervals and results in a larger proportion of statistically significant treatment effects in the simulations (24%, 13%, 13% by decreasing sample size). Recall that there is no true treatment effect and observed treatment differences are largely due to unmeasured confounding.

The mean *E*-values are provided both for all 500 simulations (labeled “All”), where non-significant results were included in the mean calculation and given an *E*-value of 1, and the subset for which statistical significance is found (labeled “Significant”). The mean *E*-values across the full 500 simulations are all small and demonstrate a lack of robustness as should be observed in this situation where there is no true treatment effect. The mean *E*-values for the point estimates are stable and driven by the large proportion of non-significant treatment effect findings. However, among the statistically significant findings, a larger mean *E*-value was seen for the smaller sample size studies as compared to the larger studies (2.00 vs. 1.57). While this may appear contradictory—that *E*-values would suggest more robustness for the smaller sample size studies—note that this is really driven by the size of the observed treatment effect. The significant studies with smaller sample sizes have a much larger treatment effects than the significant studies with larger sample sizes. Similar findings are seen with the Robustness value.

For the benchmarking exercise, we focused on the cases where the initial treatment effect analysis showed a statistically significant treatment difference. We then used benchmarking to determine the proportion of times our benchmarked variable (baseline pain severity) would suggest the finding was robust (i.e., the benchmark variable also fell in the statistically significant area of the contour plot) or not. As the truth was a null treatment effect in this simulation, ideally the benchmarking would not be robust when the treatment effect analysis was significant. Results demonstrated the added value of utilizing a structured sensitivity analysis such as proposed here given that the majority of times the treatment difference was found to be significant, the sensitivity analysis would fail to conclude this was a robust finding. For instance, with $N = 400$, 8.0% of the time a false claim of a treatment effect would have been made by the initial standard analysis but in only 2.2% of the cases this would have been viewed as a robust result. As the sample size increased to 1600, the number of times the initial treatment effect analysis found a significant effect increased, but the percentage of confirmations by the sensitivity analyses decreased to less than 1%.

Lastly, we evaluated the impact of sample size on the use of an internal data set. As in the pilot study, the internal sample (with the information on the unmeasured confounder) was 22% of the

**TABLE 3** | Simulation results: Treatment effects estimates, *E*-values, and benchmarking.

| | Sample size | | |
|---|---|---|---|
| **Measure** | **Half ($N=400$)** | **Same ($N=800$)** | **Double ($N=1600$)** |
| Treatment effect estimates (means) | | | |
| Point estimate | −0.16 | −0.13 | −0.14 |
| SD | 0.23 | 0.18 | 0.13 |
| LL | −0.76 | −0.47 | −0.39 |
| UL | 0.39 | 0.22 | 0.11 |
| Proportion significant | 0.13 | 0.13 | 0.24 |
| *E*-values | | | |
| Point estimate all; significant | 1.10; 2.00 | 1.09; 1.78 | 1.14; 1.57 |
| UL all; significant | 1.03; 1.29 | 1.03; 1.27 | 1.05; 1.19 |
| Benchmarking | | | |
| Significant and robust | 2.2% | 2.3% | 0.7% |
| Significant and not robust | 5.8% | 10.7% | 23.3% |

*Note:* Numbers represent mean values across the 500 simulations unless otherwise specified. Benchmarking results are from the Simulation Framework, other methods were similar.
Abbreviations: LL: Lower confidence limit; SD: Standard deviation; UL: Upper confidence limit.

**TABLE 4** | Simulation Results: Adjusted treatment effect estimates using the internal unmeasured confounding data and the control variable approach.

| | Sample size | | |
|---|---|---|---|
| **Measure** | **Half ($N=400$)** | **Same ($N=800$)** | **Double ($N=1600$)** |
| Adjusted treatment effect estimates | | | |
| Point Estimate | 0.01 | −0.01 | −0.02 |
| SD | 0.31 | 0.20 | 0.14 |
| LL | −0.59 | −0.41 | −0.30 |
| UL | 0.61 | 0.39 | 0.27 |
| Proportion significant | 0.08 | 0.09 | 0.07 |

*Note:* Numbers represent mean values across the 500 simulations unless otherwise specified.
Abbreviations: LL: Lower confidence limit; SD: Standard deviation; UL: Upper confidence limit.

full sample, and thus varied in proportion with the size of the full study. Multiple Imputation, the Control Variable analysis, and Bayesian Regression models were used to incorporate the internal subsample data and produce an updated treatment effect estimate in each of the 500 simulations. Table 4 provides the results of the Control Variable approach, which was representative of all methods.

Even with the smaller sample size, an internal subsample of 22% was sufficient to produce a nearly unbiased treatment effect estimate. This is consistent with the findings of Stamey et al. [27] who found that randomly sampled internal samples are effective even with small sample sizes. However, the impact on the confidence limits (much tighter for the larger samples) of the adjusted treatment effect show the benefit of a larger sample. Also, the confidence interval widths, compared to Table 3, show a slightly greater uncertainty in our estimates after incorporating the information on unmeasured confounding from the subsample. Thus, at the design stage of a study one could vary both the size of the study and internal sample via simulations to better plan for an appropriate study size that would produce an adjusted estimate with desired confidence interval widths allowing detection of the expected treatment effect. Clearly, more tools for researchers to simplify this assessment are needed. As a reminder, simply increasing the size of the main study will not reduce the size of the bias on the treatment effect estimate—so careful consideration should be given to other data in addition to the study sample size.

## 4 | Section 4—Conclusions

Comparative observational studies should include pre-planned and quantitative assessment of the potential impact of unmeasured confounding. As reliance on such RWE by decision makers grows, so should consistent use of quality sensitivity analyses surrounding core statistical assumptions.

Here we proposed a structured approach to guide the evaluation of potential bias due to unmeasured confounding both at the stage when developing a protocol and at the data analysis stage. In each stage we present 3 questions that help researchers address the potential impact of unmeasured confounding along with a toolbox of methods and links to software for implementation. The goal at the design stage is to help forecast whether the planned study is likely to produce a causal treatment effect estimate robust against expected levels of bias due to unmeasured confounding and to stimulate planning for additional data collection if needed. At the analysis stage, the goal is to provide quantitative assessment of the robustness of the observed finding to potential unmeasured confounding, using all available information. We believe such an approach will provide greater information regarding the robustness of comparative observational research to medical decision makers.

We demonstrated the use of this good practice guidance with the data simulated from an observational study (REFLECTIONS). While commonly used analyses found a statistically significant treatment effect, sensitivity analyses showed that the results were not robust to potential unmeasured confounding at the strength of the largest measured confounder (baseline pain severity). Incorporating additional information on an identified unmeasured confounder enabled more accurate benchmarking and adjusted treatment effect estimates close to the true treatment effect.

The simulation study also assessed the impact of study sample size on the operating characteristics of our sensitivity analysis guidance using the setup for the simulated REFLECTIONS data. Results demonstrated that the benchmarking and internal data approaches from the guidance would have consistently suggested lack of robustness of any estimated treatment effect. Use of the guidance led to the correct inferences in this simulation, since there was no true treatment effect. Simulations involving the use of additional data produced near unbiased treatment effect estimates even with half the sample size of our pilot study. The simulations also highlighted that more tools are needed to help researchers at the design stages to plan for sensitivity analyses, such as generating simulations to determine how much internal or external data are necessary to produce a robust finding in the presence of some level of unmeasured confounding. Of course, increasing the sample size will not erase issues with bias.

We note that this work is not without limitations. First, we offer a single application; greater use of this guidance will likely bring refinement. For instance, the internal Supporting Information was beneficial in our example but may not be sufficient in every situation. Further, reliable information on confounding in even a small subsample will not always be available outside of simulated data settings. Our toolbox will need to change over time as additional methods and software options become available. We did not demonstrate all potentially useful methods, such as those that employ identification strategies that side-step the requirement of "no unmeasured confounders" in favor of other demanding assumptions (e.g., negative controls or instrumental variables analyses). Our simulation study evaluated the impact of sample size on use of the guidance but was not a thorough evaluation of other factors such as effect size and level of confounding. Other types of research questions, such as those addressed in non-inferiority studies and the use of real-world controls for clinical trials, have not been addressed here. While not directly applicable and would require modification (such as addressing database ignorability in the use of real-world controls) [48], we believe the principles applied here may prove useful for future research for settings such as real-world controls. Unmeasured confounding is just one of several assumptions required for causal inference and we recommend a careful evaluation of the validity of all assumptions. We focused on unmeasured confounding given its potential to cause significant bias and the under-utilization of sensitivity analyses for this assumption—though we acknowledge that focusing on a single type of bias potentially results in overconfidence in any robustness findings. Lastly, in the DAGs we have oversimplified by considered each potential confounder as measured or not, while in practice some may be measured with substantial error or have substantial missing data [37].

In summary, consistent application of quantitative sensitivity analyses will provide decision makers with more reliable information on the robustness of RWE and will lead to greater acceptance of quality RWE and ultimately better patient outcomes. Our hope is that this work will accelerate the growing trend toward consistent and high-quality application of quantitative sensitivity analyses for unmeasured confounding.

## Conflicts of Interest

D.F., K.S., and M.S. are full-time employees and stockholders of Eli Lilly & Company. X.Z. is a full-time employee and stockholder of CSL Behring. C.H., J.S., N.D., P.D., and S.Y. have been paid consultants by Eli Lilly and Company for the project of sensitivity analysis against unmeasured confounding.

## Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

## References

1. I. J. Dahabreh and K. Bibbins-Domingo, "Causal Inference About the Effects of Interventions From Observational Studies in Medical Journals," *JAMA* 331 (2024): E1–E9.

2. E. Patorno, S. Schneeweiss, C. Gopalakrishnan, D. Martin, and J. M. Franklin, "Using Real-World Data to Predict Findings of an Ongoing Phase IV Cardiovascular Outcome Trial: Cardiovascular Safety of Linagliptin Versus Glimepiride," *Diabetes Care* 42, no. 12 (2019): 2204–2210.

3. M. Levenson, W. He, J. Chen, Y. Fang, D. Faries, and B. A. Goldstein, "Biostatistical Considerations When Using RWD and RWE in Clinical Studies for Regulatory Purposes: A Landscape Assessment," *Statistics in Biopharmaceutical Research* 15, no. 1 (2023): 3–13.

4. C. V. Arianth and E. F. Schisterman, "Hidden Biases in Observational Epidemiology: The Case of Unmeasured Confounding," *BJOG* 125, no. 6 (2018): 644–646.

5. T. J. VanderWeele and O. A. Arah, "Unmeasured Confounding for General Outcomes, Treatments, and Confounders: Bias Formulas for Sensitivity Analysis," *Epidemiology* 22, no. 1 (2011): 42–52.

6. X. Zhang, D. E. Faries, N. Boytsov, J. D. Stamey, and J. W. Seaman, Jr., "A Bayesian Sensitivity Analysis to Evaluate the Impact of Unmeasured Confounding With External Data: A Real World Comparative Effectiveness Study in Osteoporosis," *Pharmacoepidemiology and Drug Safety* 25, no. 9 (2016): 982–992.

7. M. L. Berger, H. Sox, R. J. Willke, 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," *Value in Health* 20, no. 8 (2017): 1003–1008.

8. M. R. Blum, Y. J. Tan, and J. P. A. Ioannidis, "Use of E-Values for Addressing Confounding in Observational Studies—An Empirical Assessment of the Literature," *International Journal of Epidemiology* 49 (2020): 1482–1494.

9. N. A. Dreyer, S. Schneeweiss, B. McNeil, et al., "GRACE Principles: Recognizing High-Quality Observational Studies of Comparative Effectiveness," *American Journal of Managed Care* 16, no. 6 (2010): 467–471.

10. M. Berger, B. C. Martin, D. Husereau, et al., "A Questionnaire to Assess the Relevance and Credibility of Observational Studies to Inform Healthcare Decision Making: An ISPOR-AMCP- NPC Good Practice Task Force," *Value in Health* 17, no. 2 (2014): 143–156.

11. N. A. Dreyer, A. Bryant, and P. Velentgas, "The GRACE Checklist: A Validated Assessment Tool for High Quality Observational Studies of Comparative Effectiveness," *Journal of Managed Care & Specialty Pharmacy* 22, no. 10 (2016): 1107–1113.

12. P. Velentgas, N. A. Dreyer, P. Nourjah, S. R. Smith, and M. M. Torchia, "Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. Conducted Under Contract No. 290-2005-0035-I. AHRQ Publication No. 12(13)-EHC099. Rockville, MD: Agency for Healthcare Research and Quality," (2013).

13. M. J. Uddin, R. H. H. Groenwold, M. S. Ali, et al., "Methods to Control for Unmeasured Confounding in Pharmacoepidemiology: An Overview," *International Journal of Clinical Pharmacy* 38, no. 3 (2016): 1–10.

14. A. J. Streeter, N. X. Lin, L. Crathorne, et al., "Adjusting for Unmeasured Confounding in Non-randomised Longitudinal Studies: A Methodological Review," *Journal of Clinical Epidemiology* 87 (2017): 23–34.

15. X. Zhang, D. E. Faries, H. Li, J. D. Stamey, and G. W. Imbens, "Addressing Unmeasured Confounding in Comparative Observational Research," *Pharmacoepidemiology and Drug Safety* 27 (2018): 373–382.

16. T. L. Lash, M. P. Fox, and A. K. Fink, *Applying Quantitative Bias Analysis to Epidemiologic Data* (New York: Springer Science + Business Media, 2009), 194.

17. T. J. VanderWeele and P. Ding, "Sensitivity Analysis in Observational Research: Introducing the E-Value," *Annals of Internal Medicine* 167, no. 4 (2017): 268–274.

18. X. Zhang, J. Stamey, and M. B. Mather, "Assessing the Impact of Unmeasured Confounders for Credible and Reliable Real-World Evidence," *Pharmacoepidemiology and Drug Safety* 2020, no. 29 (2020): 1219–1227.

19. C. Cinelli and C. Hazlett, "Making Sense of Sensitivity: Extending Omitted Variable Bias," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 82 (2020): 39–67.

20. T. J. VanderWeele and M. B. Mathur, "Commentary: Developing Best-Practice Guidelines for the Reporting of E-Values," *International Journal of Epidemiology* 49, no. 5 (2020): 1495–1497.

21. R. L. Robinson, K. Kroenke, P. Mease, et al., "Burden of Illness and Treatment Patterns for Patients With Fibromyalgia," *Pain Medicine* 13 (2012): 1366–1376.

22. T. J. VanderWeele, P. Ding, and M. B. Mathur, "Technical Considerations in the Use of the E-Value," *Journal of Causal Inference* 7, no. 2 (2019): 1–11.

23. M. B. Mathur, L. H. Smith, P. Ding, and T. J. Van der Weele, "Evalue Package," (2021), https://cran.r-project.org/web/packages/EValue/index.html.

24. A. Linden, M. B. Mathur, and T. J. VanderWeele, "Conducting Sensitivity Analyses for Unmeasured Confounding in Observational Studies Using E-Values: The Evalue Package," *Stata Journal* 20, no. 1 (2020): 162–175.

25. V. Dorie, M. Harada, N. B. Carnegie, and J. Hill, "A Flexible, Interpretable Framework for Assessing Sensitivity to Unmeasured Confounding," *Statistics in Medicine* 35 (2016): 3453–3470.

26. N. B. Carnegie, M. Harada, and J. L. Hill, "Assessing Sensitivity to Unmeasured Confounding Using a Simulated Potential Confounder," *Journal of Research on Educational Effectiveness* 9, no. 3 (2016): 395–420.

27. J. Stamey, D. Beavers, D. Faries, K. L. Price, and J. W. Seaman, Jr., "Bayesian Modeling of Cost-Effectiveness Studies With Unmeasured Confounding: A Simulation Study," *Pharmaceutical Statistics* 13 (2014): 94–100.

28. D. Faries, X. Peng, M. Pawaskar, K. Price, J. Stamey, and J. Seaman, "Evaluating the Impact of Unmeasured Confounding With Internal Validation Data: An Example Cost Evaluation in Type 2 Diabetes," *Value in Health* 16 (2013): 259–266.

29. E. J. Bedrick, R. Christensen, and W. Johnson, "A New Perspective on Priors for Generalized Linear Models," *Journal of the American Statistical Association* 91, no. 436 (1996): 1450–1460.

30. R. Hebdon, J. Stamey, D. Kahle, and X. Zhang, "Unmconf: Modeling With Unmeasured Confounding," https://cran.r-project.org/web/packages/unmconf/index.html.

31. S. Yang and P. Ding, "Combining Multiple Observational Data Sources to Estimate Causal Effects," *Journal of the American Statistical Association* 115 (2020): 1540–1554.

32. C. J. Girman, D. Faries, P. Ryan, M. Rotelli, M. Belger, and B. Binkowitz, "O'Neill R, for the Drug Information Association CER Working Group. Pre-Study Feasibility and Identifying Sensitivity Analyses for Protocol Pre-Specification in Comparative Effectiveness Research," *Journal of Comparative Effectiveness* 3, no. 3 (2014): 259–270.

33. P. W. G. Tennant, E. J. Murray, K. F. Arnold, et al., "Use of Directed Acyclic Graphs (DAGs) to Identify Confounders in Applied Health Research: Review and Recommendations," *International Journal of Epidemiology* 50, no. 2 (2021): 620–632.

34. J. Textor, B. van der Zander, M. K. Gilthorpe, M. Liskiewicz, and G. T. H. Ellison, "Robust Causal Inference Using Directed Acyclic Graphs: The R Package 'Dagitty'," *International Journal of Epidemiology* 45, no. 6 (2016): 1887–1894.

35. J. C. Digitale, J. N. Martin, and M. M. Glymour, "Tutorial on Directed Acyclic Graphs," *Journal of Clinical Epidemiology* 142 (2021): 264–267.

36. K. D. Ferguson, M. McCann, S. V. Katikireddi, et al., "Evidence Synthesis for Construction Directed Acyclic Graphs (ESC-DAGs): A Novel and Systematic Method for Building Directed Acyclic Graphs," *International Journal of Epidemiology* 49, no. 1 (2020): 322–329.

37. F. Kuehne, M. Arvandi, L. M. Hess, et al., "Assessing the Impact of Biases When Analyzing Real World Data: The Case of 2nd Line Chemotherapy in Ovarian Cancer Women," *Journal of Clinical Epidemiology* 152 (2022): 269–280.

38. Y. Fang, W. He, X. Hu, and H. Wang, "A Method for Sample Size Calculation via E-Value in the Planning of Observational Studies," *Pharmaceutical Statistics* 20, no. 1 (2021): 163–174.

39. S. Schneeweiss, "Sensitivity Analysis and External Adjustment for Unmeasured Confounders in Epidemiologic Database Studies of Therapeutics," *Pharmacoepidemiology and Drug Safety* 15 (2006): 291–303.

40. Z. Zhang, J. Uddin, J. Cheng, and T. Huang, "Instrumental Variable Analysis in the Presence of Unmeasured Confounding," *Annals of Translational Medicine* 6, no. 10 (2018): 182.

41. M. J. Schuemie, P. B. Ryan, W. DuMouchel, M. A. Suchard, and D. Madigan, "Interpreting Observational Studies: Why Empirical

Calibration Is Needed to Correct p-Values," *Statistics in Medicine* 33, no. 2 (2014): 209–218.

42. D. W. Flanders, L. A. Waller, Q. Zhang, D. Getahun, M. Silverberg, and M. Goodman, "Negative Control Exposures: Causal Effect Identifiability and Use in Probabilistic Bias and Bayesian Analyses With Unmeasured Confounders," *Epidemiology* 33, no. 6 (2022): 832–839.

43. D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys* (New York: John Wiley & Sons, 1987), 258.

44. X. Peng, R. L. Robinson, P. Mease, et al., "Long-Term Evaluation of Opioid Treatment in Fibromyalgia," *Clinical Journal of Pain* 31 (2015): 7–13.

45. M. Ho, M. van der Laan, H. Lee, et al., "The Current Landscape in Causal Inference Frameworks for Design and Analysis of Studies Using Real-World Data and Evidence," *Statistics in Biopharmaceutical Research* 15, no. 1 (2023): 29–42.

46. S. Gruber and M. van der Laan, "Tmle: An R-Package for Targeted Maximum Likelihood Estimation," *Journal of Statistical Software* 51 (2012): 13.

47. M. P. Fox, R. F. MacLehose, and T. L. Lash, *Applying Quantitative Bias Analysis to Epidemiologic Data* (New York: Springer Science + Business Media, 2021), 467.

48. M. Shan, D. Faries, A. Dang, X. Zhang, Z. Cui, and K. Sheffield, "A Simulation-Based Evaluation of Statistical Methods for Hybrid Real World Control Arms in Clinical Trials," *Statistics in Biosciences* 14 (2022): 259–284.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section.