Quantitative bias analysis in practice: Review of software for regression with unmeasured confounding

E Kawabata[1,2], K Tilling[1,2], RHH Groenwold[3,4], and RA Hughes[1,2]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom
[2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[3]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
[4]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Abstract

Failure to appropriately account for unmeasured confounding in analyses may lead to bias and erroneous conclusions. Quantitative bias analysis (QBA) for unmeasured confounding is used to quantify the potential direction and impact of the bias. The adoption of QBA by applied researchers has been slow, partly due to the focus of methods for binary outcomes and exposure, and partly due to the lack of accessible software. We provide a review of the latest developments in QBA software during 2010 to 2020. We describe in detail 5 QBA methods and their software implementations that can be applied when the analysis of interest is a linear regression. We illustrate application of these software programs to real data and provide R and Stata software code along with a practice example. In our review, all software implementations were of deterministic QBA methods and mostly implemented as R packages. Graphical presentations of the results and benchmarking are useful aids of interpretation. However, sole reliance on summary measures at tipping points should be discouraged as it can encourage analysts to place too much emphasis on statistical significance. The diversity of QBA methods presents challenges in the widespread uptake of QBA. Guidelines are needed on the appropriate choice of QBA method, along with provision of software implementations in platforms other than R.

Keywords: Causal inference; Linear regression; Review; Sensitivity analysis; Software; Unmeasured confounding.

1 Introduction

The main aim of many epidemiology studies is to estimate the causal effect of an exposure on an outcome (here onward, shortened to exposure effect). In observational studies participants are not randomised to exposure (or treatment) groups. Consequently, factors that affect the outcome are typically unevenly distributed among the exposure groups, and a direct comparison between the exposure groups will likely be biased due to confounding. Standard adjustment methods (such as standardization, inverse probability weighting, regression adjustment, g-estimation, stratification and matching) assume all of these confounders are measured without error; that is, no unmeasured (or residual) confounding [1]. Failure to appropriately account for unmeasured or poorly measured confounders in analyses may lead to invalid inference (e.g., [2–4]).

There are several approaches which can account for unmeasured confounding at the analysis stage, including instrumental variable analysis, negative controls, perturbation variable analysis, methods that use confounder data collected on a study sub-sample (e.g., propensity score calibration analysis), and quantitative bias analysis (QBA; also known as a sensitivity analysis) [5]. A QBA is applied when a study cannot control for unmeasured confounding using available data (e.g., lacks an appropriate instrument or sub-sample data on the unmeasured confounders). It quantifies the potential impact of unmeasured confounding on an estimate of the exposure effect and assesses whether the conclusions of the study change under different assumptions about the unmeasured confounding.

Lack of knowledge about QBA, and of analyst-friendly methods and software have been identified as barriers to the widespread implementation of a QBA [6–8]. In the past decade, there have been several reviews of QBA methods [2, 5, 8–12]. Only one of these, [10], reviewed software implementations and since its publication in 2014 there have been many new software implementations. Also, comparisons of QBA methods have primarily been limited to analyses with a binary outcome [9, 13–20].

Our paper provides an up to date review on available software for researchers wishing to implement a QBA to address unmeasured confounding in studies that consider the total effect of a single exposure. We then describe, illustrate and compare QBA methods, and their software, applicable when the analysis of interest is a linear regression. We illustrate how to apply these methods using two real-data examples: the $2015 - 2016$ National Health and Nutrition Examination Survey (NHANES) study [21] and the Barry Caerphilly Growth (BCG) study [22, 23].

2 Quantitative bias analysis for unmeasured confounding

We want to estimate the effect of an exposure (or treatment) X on an outcome Y. The $Y - X$ association is confounded by measured covariates C and unmeasured confounders U. The naive estimate of the exposure effect, $\hat{\beta}_{X|C}$, assumes no

unmeasured confounding and is estimated by controlling for C only.

We can use a QBA to quantify the likely magnitude and direction of the bias, due to unmeasured confounding, under different plausible assumptions about U. Generally, a QBA requires a model (known as a bias model) for the observed data, Y, X and C, and unmeasured data, U. The bias model will include one or more parameters (known as bias or sensitivity parameters) which cannot be identified from the observed data. For example, bias parameters that specify the strength of the association between U and X, and between U and Y given X [15]. Information about the likely values of these bias parameters may be obtained from external sources (such as external validation studies, published literature, or expert opinion) [7], and from benchmarking (also known as calibration) where strengths of associations of measured covariates C with X and Y are used as benchmarks [24]. We shall denote the bias parameters by ϕ and the bias-adjusted estimate of the exposure effect assuming ϕ by $\hat{\beta}_{X|C,U(\phi)}$.

A QBA is often conducted as a tipping point analysis, where the analyst identifies the values of ϕ that correspond to a change in the study conclusions (known as the "tipping point"). A tipping point analysis may be applied to the point estimate or confidence interval of the exposure effect; for example, to identify the values of ϕ corresponding to a null effect, or the values of ϕ corresponding to a statistically insignificant effect (of a non-null point estimate). If the values of ϕ at the tipping point(s) are considered unlikely then the study conclusions are said to be robust to unmeasured confounding.

There are two broad classes of QBA methods: deterministic and probabilistic. A deterministic QBA specifies a range of values for each bias parameter of ϕ and then calculates $\hat{\beta}_{X|C,U(\phi)}$ for all combinations of the specified values of ϕ. Typically, the results are displayed as a plot (or table) of $\hat{\beta}_{X|C,U(\phi)}$ against different values of ϕ. Unlike a deterministic QBA, a probabilistic QBA explicitly models the analyst's assumptions about which combinations of ϕ are most likely to occur and incorporates their uncertainty about ϕ [6,25]. A probabilistic QBA achieves this by specifying a prior probability distribution for ϕ [6]. Averaging over this probability distribution generates a distribution of estimates $\hat{\beta}_{X|C,U(\phi)}$ which is summarised to give a point estimate (i.e., the most likely $\hat{\beta}_{X|C,U(\phi)}$ under the QBA's assumptions) and an interval estimate (i.e., defined to contain the true exposure effect with a pre-specified probability) which accounts for uncertainty due to sampling variability, unmeasured confounding and about the true values of ϕ.

3 Overview of available software

The aim of the literature search was to give a brief overview of the available software implementations of QBA published between 1st January 2010 and 31st December 2020 (inclusive). We have focused on unmeasured confounding for a main effects analysis, where the exposure effect is quantified using a risk difference, mean difference, risk ratio, odds ratio, and hazard ratio. We have not covered the special case of unmeasured confounding for mediation analysis, or multilevel settings. Also, we

focus on "software programs" (i.e., packages or commands) that the analyst can apply to their data without adapting the source code.

Our literature search was conducted in three stages. In stage 1, we used Web of Science to identify papers that mentioned "quantitative bias analysis" and "unmeasured confounding" (or their synonyms) in either the title, abstract or as keywords (see Supplementary Box 1 for our search strategy). In stage 2, the abstracts were reviewed by two independent reviewers to determine if they were eligible for stage 3 (data extraction), with any disagreements resolved by consensus.

Eligible abstracts were articles or reviews published in a journal which either introduced a new QBA method or software implementation, or was a comparison or review of existing QBA methodology, or a tutorial-style paper on how to conduct a QBA. Examples of ineligible abstracts were meeting abstracts, commentaries, and articles where either the applied analysis was the primary focus (and so included limited information on the statistical methodology used) or the authors did not conduct a QBA (e.g., only mentioned QBA as further work). In stage 3, we read the full text to check its eligibility and extracted information about the analysis of interest (e.g., type of outcome and exposure variables), the QBA method (e.g., requires individual participant or summary data) and the software (e.g., generates graphical output).

After excluding duplicates, our Web of Science search identified 301 papers. We excluded 224 and 63 at the second and third stages respectively, leaving 14 papers for data extraction.

Table 1 summarises the main features of the 14 software programs we identified, five of which implement a QBA for an observational study with matched pairs or matched sets containing multiple controls. Twelve programs are implemented in software environment R [26], with three of them also available in Stata [27] and as a Shiny application [28]. Out of the 14 programs, three require the exposure to be binary, and eight can only be applied to one type of outcome variable. A program's range of applicability will depend on its underlying QBA method. For example, programs *sensitivitymv*, *sensitivitymw* and *submax* all implement a QBA method which computes sensitivity bounds for P-values, and so can be applied to many types of outcome variable. All programs can be applied to individual participant data, with three programs also applicable for summary data (e.g., exposure estimates and standard errors). Furthermore, programs *E-value* and *konfound* can compute a QBA for a single study and for a meta-analysis. Not all packages produced a graphical output or reported benchmark statistics. All programs allow the analysis of interest to adjust for measured covariates C.

All programs implement a deterministic QBA and can be applied as a tipping point analysis. Note that, program *ui* reports an uncertainty interval which is defined be the union of all confidence intervals over the specified values of ϕ [29]. Other specialisations include: (i) *causalsens*, *treatSens* and *ui* are applicable when the estimand of interest is the average exposure effect among the exposed (or unexposed), (ii) *tukeySens* is applicable for quantile exposure effects, and (iii) *treatSens* and *tukeySens* allow for flexible modelling of continuous outcome Y.

Table 1 Software programs implementing quantitative bias analysis for unmeasured confounding, published between 2010 and 2020

	Applicable analysis of interest				
Program (environment)	Type of analysis	Outcome	Exposure	Bench-marking	Graphical output
causalsens (R) [30]	simple[a]	con[b]	bin[c]	yes	yes
E-Value[d] (R, Stata, Shiny) [31–34]	simple, meta-analysis	bin, con, TTE[e]	bin, con	no	yes
gsa (Stata) [10, 35]	simple	bin, con	bin, con, cat[f]	yes	yes
isa (Stata) [36]	simple	con	bin	yes	yes
konfound (R, Stata, Shiny) [37]	simple, meta-analysis	bin, con	bin, con	yes	yes
sensemakr (R, Stata, Shiny) [38, 39]	simple	con	bin	yes	yes
sensitivityCalibration (R) [24]	matched	con	bin	yes	yes
sensitivityCaseControl (R) [40]	matched	bin	bin	no	no
sensitivitymv (R) [41]	matched	all[g]	bin	no	no
sensitivitymw (R) [42]	matched	all	bin	no	no
submax (R) [43]	matched	all	bin	no	no
treatSens (R) [44, 45]	simple	con	bin, con	yes	yes
tukeySens (R) [46]	simple	con	bin	yes	yes
ui (R) [47]	simple	bin	bin	no	yes

[a]Unmatched analysis from a single study; [b]continuous variable; [c]binary variable; [d]R package EValue, Stata command evalue and Shiny web application E-value; [e]time to event variable; [f]categorical variable; [g]includes Huber-Maritz M-scores, ranked scores, outcomes compatible with permutational t-test.

4 Quantitative bias analysis methods for linear regression

We selected the following five QBA programs to describe and illustrate in this paper: *treatSens* [44, 45], *causalsens* [30], *sensemakr* [48], *E-value* [31], and *konfound* [37]. We selected software programs from Table 1 applicable for an unmatched analysis, where the exposure effect is estimated by a linear regression model. We decided to focus on relatively straightforward methods and so excluded program *tukeySens* which accommodates flexible models of complex data (e.g., Dirichlet processes mixture models). Also, for reasons of brevity, we excluded programs *isa* and *gsa* as they are similar to the more recently published *treatSens*.

Below, we summarise the QBA methods implemented by the selected programs (see the Supplementary Materials for detailed descriptions, including software details). The main features of the selected QBA methods and software (not given in Table 1) are shown in Table 2 and their strengths and weaknesses in Table 3.

4.1 *treatSens*

R package *treatSens* simulates U using its bias model and then estimates $\hat{\beta}_{X|C,U(\phi)}$ from a linear regression of Y on X adjusted for C and the simulated U (the analysis model). The bias model consists of three sub-models: the analysis model, the treatment model which is a regression of X on C and U (e.g., linear or probit regression for continuous or binary X, respectively), and a marginal model for U (standard Normal or Bernoulli). The bias model has two bias parameters $\phi = (\zeta^Y, \zeta^Z)$: ζ^Y is the coefficient for the $Y - U$ association in the analysis model and ζ^Z is the coefficient for the $X - U$ association in the treatment model. To allow for unmeasured confounding bias in both directions (i.e., increased exposure effect, and reduced or reversed exposure effect), positive and negative values are specified for ζ^Z. By default, *treatSens* selects the range of values for ζ^Y and ζ^Z based on the residual variances of the analysis and treatment models, respectively. Additionally, *treatSens* allows analysts to specify their own ranges for ζ^Y and ζ^Z. The remaining parameters of the bias model are estimated from the observed data. To gauge the plausible magnitudes of ζ^Y and ζ^Z, the coefficients of measured covariates C (from the regressions of Y on X and C, and X on C) are used as benchmark values. All continuous variables are standardised to facilitate comparison between these benchmark values and the bias parameters.

treatSens outputs a contour plot (and tables) displaying estimates $\hat{\beta}_{X|C,U(\phi)}$ for the prespecified values of ζ^Y and ζ^Z, indicating the combinations that correspond to the tipping points for the point estimate and CI of the exposure effect. The default tipping point for the CI is 5% statistical insignificance (i.e., 95% CI for $\hat{\beta}_{X|C,U(\phi)}$ includes the null) which the analyst can change, while the tipping point for the point estimate is fixed at the null effect (i.e., $\hat{\beta}_{X|C,U(\phi)} = 0$). To help quicken the runtime of *treatSens* there is an option to specify multiple central processing unit cores for parallel processing.

4.2 *causalsens*

R package *causalsens* uses a bias model to generate a modified outcome, Y_{ϕ}^{adj}, which is adjusted for the unmeasured confounding for fixed values of ϕ. The naive analysis is then reapplied using Y_{ϕ}^{adj} instead of Y, and $\hat{\beta}_{X|C,U(\phi)}$ and its CI are the corresponding exposure effect estimates. The bias model consists of a user-specified function, called the "confounding function", which quantifies the unmeasured confounding, and a treatment model which is a logistic regression used to estimate the probability of being in the exposed group given covariates C. Note, *causalsens* requires a binary X. The confounding function [49, 50] is based on the potential outcomes framework [51]. For binary X, the confounding function quantifies the average difference in potential outcomes to exposure (or non-exposure) between the exposed and unexposed groups, where any non-zero difference is attributed to unmeasured confounding. *causalsens* supplies two choices for the confounding function, the one-sided function and the alignment function, and also allows the analyst to specify their own function. Both supplied functions are parameterised by a single bias parameter, $\phi = (\alpha)$. The one-sided function assumes the true exposure effect is identical in the exposed and unexposed groups. When $\alpha > 0$ the mean of the potential outcomes to exposure (and non-exposure) is higher for the exposed group than the unexposed group, leading $\hat{\beta}_{X|C}$ to be positively biased; and vice versa for $\alpha < 0$. (See the Supplementary Materials for details on the alignment function.) By default, *causalsens* selects 11 values for α covering the interquartile range of outcome Y. Both negative and positive values of α are selected to allow for the effect of unmeasured confounding in both directions. The analyst can also specify their own values for α.

causalsens outputs a line plot displaying estimates $\hat{\beta}_{X|C,U(\phi)}$ and its 95% CI for the prespecified values of α. The tipping point for the CI is fixed at 5% statistical insignificance. The tipping point of the point estimate is not set to a particular value as the analyst can select their own value from the vertical axis of the plot. Since α is difficult to interpret, *causalsens* offers an alternative parameterisation $R^2_{\alpha} = sgn(\alpha) \times R^2_U$, where $sgn(\alpha)$ denotes the sign of α (i.e., direction of bias) and R^2_U is the partial R^2 for the proportion of residual variance in the potential outcomes explained by U. For this alternative parameterisation, the line plot indicates benchmarks for α based on the partial R^2 values of the measured covariates C from the naive analysis (i.e., for each measured covariate C_j, the benchmark is $R^2_{Y\sim C_j|X,C_{-j}}$, the proportion of the variance of Y, not explained by X and the remaining covariates C_{-j}, that is explained by C_j).

4.3 *sensemakr*

sensemakr expresses the magnitude of the bias as a function of estimated quantities from the naive analysis and bias parameters ϕ. For a user-specified value of ϕ, *sensemakr* calculates the magnitude of the bias, which is then used to derive $\hat{\beta}_{X|C,U(\phi)}$ and its corresponding standard error. There are two bias parameters $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$ which denote the partial R^2 of U with the exposure and

outcome, respectively. By default, *sensemakr* sets the range of values for $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$ and the direction of the effect of U is set to reduce the absolute value of the exposure effect. The analyst can override these default settings. *sensemakr* calculates upper bounds (called 'benchmark bounds') for bias parameters $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$ based on the measured covariates. Importantly, these benchmarks are based on the intended analysis, $Y|X, C, U$ and are not solely derived from the naive analysis, $Y|X, C$ [48].

sensmakr outputs a table of benchmark bounds and two contours plots of estimates $\hat{\beta}_{X|C,U(\phi)}$ and corresponding t-value for the prespecified values of $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$, indicating the combinations that correspond to the tipping points for the point estimate and CI of the exposure effect. The default tipping points are the null effect and 5% statistical insignificance, both of which can be changed by the analyst.

4.4 *E-value*

Program *E-value* reports a single summary measure, called an E-value, which quantifies the minimum magnitude of the associations between U and X and between U and Y, conditional on C, needed to move $\hat{\beta}_{X|C}$ to a specified value (such as the null) or render $\hat{\beta}_{X|C}$ to be statistically insignificant. The E-value is a positive number ≥ 1 with higher values indicating that greater levels of unmeasured confounding (i.e., stronger $X-U$ and $Y-U$ associations) are required to change the study conclusions (e.g., reduce the exposure effect to the null). The rationale of the E-value is based on an upper bound of a bias factor, BF_ϕ, for a given level of unmeasured confounding ϕ, where BF_ϕ is used to derive bounds for the bias-adjusted results. This upper bound, BF_ϕ, is expressed as a function of estimated quantities from the naive analysis and two bias parameters RR_{XU} and RR_{UY} which represent the strength of the $X-U$ and $Y-U$ associations on the risk ratio scale, respectively (see the Supplementary Material for more details). Setting these two bias parameters as equal, ϕ_{equal}, the E-value is the minimum value of ϕ_{equal} at which BF_ϕ equals a set tipping point.

The E-value is defined (and interpreted) on the risk ratio scale. However, program *E-value* can also calculate an E-value for a mean or risk difference, or odds or hazard ratio by first converting the effect measure to a risk ratio (under various assumptions [31]). The program outputs the E-value for the point estimate and CI limit, along with a line graph depicting the combinations of RR_{XU} and RR_{UY} that result in the tipping points values. By default, the tipping points for the point estimate and statistical significance are the null effect and 5%, respectively, both of which can be changed by the analyst.

Note that, the E-value is a measure of sensitivity to unmeasured confounding for an extreme scenario since the prevalence of X among those without U is assumed to be at a value that generates maximum bias [15]. For example, when the tipping point is a null exposure effect, then the prevalence of X among those without U is 0% (i.e., all exposed people have a non-zero value for U). Therefore, program *E-value* does not calculate benchmark values for bias parameters RR_{UY} and RR_{XU}.

Instead, for purposes of comparison, VanderWeele and Ding suggest omitting each measured covariate in turn and recalculating the E-value [31].

4.5 *konfound*

konfound assesses sensitivity to a change in the statistical significance/insignificance status of $\hat{\beta}_{X|C}$. This includes the scenario where U explains all of the statistical significance of $\hat{\beta}_{X|C}$ (i.e., $\hat{\beta}_{X|C}$ is statistically significant but $\hat{\beta}_{X|C,U(\phi)}$ is statistically insignificant) and the converse scenario where U restores the statistical significance of $\hat{\beta}_{X|C}$ (i.e., $\hat{\beta}_{X|C}$ is statistically insignificant but $\hat{\beta}_{X|C,U(\phi)}$ is statistically significant). *konfound* refers to the first scenario as U "invalidating inference" and the second as U "sustaining inference". *konfound* reports two measures that quantify the level of unmeasured confounding necessary to change conclusions on statistical significance: percent bias and impact threshold. Percent bias is a measure of the minimum percentage of $\hat{\beta}_{X|C}$ that would need to be explained away by U in order for unmeasured confounding to invalidate inference. Impact threshold is a measure of the minimum strength of the partial correlation between U and Y, and U and X (conditional on C) in order for unmeasured confounding to invalidate or sustain inference. For both measures, larger (absolute) values indicate greater robustness to unmeasured confounding.

konfound outputs the percent bias, depicted by a bar graph (called a "threshold plot") and the impact threshold, depicted by a causal-type diagram (called a "correlation plot"). Also, *konfound* outputs a table of benchmarks for the impact threshold which are based on partial correlations of the measured covariates with X and Y from the naive analysis (i.e., the product of the partial correlations between measured covariate C_j and X, and between C_j and Y (conditional on the remaining confounders C_{-j}). By default, the significance level is 5% and the null hypothesis is "no exposure effect", both of which can be changed by the analyst.

Table 2 Features of the five quantitative bias analysis methods and software implementations

Feature	*treatSens*	*causalsens*	*sensemakr*	evalue	*konfound*
Effect measure	MD[a]	MD	MD	RR[b], RD[c], OR[d], HR[e], MD	MD, OR
Estimand	ATE[f], ATT[g]	ATE, ATT	not specified	not specified	not specified
Variable type of U	binary, continuous	unrestricted	unrestricted	unrestricted	unrestricted
$U \perp\!\!\!\perp C$	Yes	Yes	Yes	Yes	Yes
U can modify $Y - X$ association	No	Not specified	No	Yes	Not specified
Multiple confounders, U represents	linear combination	unrestricted	linear combination	unrestricted	linear combination
Number of bias parameters	2	1	2	2	2
Tipping point for					
exposure effect	null only	null only	user set	user set	not applicable
statistical significance	user set	5%	user set	user set	user set
Plot types	contour line	line with CI[h]	contour line	line	bar, causal diagram

a: mean difference; b: risk ratio; c: risk difference; d: odds ratio; e: hazard ratio; f: average treatment effect; g: average treatment effect among treated; h: confidence interval.

Table 3 Strengths and weaknesses of five quantitative bias analysis (QBA) methods assessing sensitivity of the effect of exposure X on outcome Y (given measured covariates C) to unmeasured confounding by U.

Method	Strengths	Weaknesses
treatSens	Applicable when $Y\|X, C$ is a Bayesian additive regression tree	Nontrivial runtimes increasing with sample size
	QBA method familiar to users of multiple imputation	Conservative for multiple confounders
causalsens	User-specified confounding function provides flexibility	Limited control over $Y - U$ and $X - U$ associations as both encapsulated by single parameter
	Few restrictions placed on functional form of U	Software does not allow the user to change the tipping points
sensemakr	Able to group multiple measured covariates for benchmarking	Conservative for multiple confounders
	Benchmarking accounts for U	
E-value	Applicable to wide range of effect measures	Approximation to risk ratio scale requires additional assumptions
	Requires only summary data; adapted for meta analyses	Only reports results at tipping point
		Measure of sensitivity for extreme prevalence of the confounder
konfound	Requires only summary data; adapted for meta analyses	Only considers sensitivity of statistical significance
		Only reports results at tipping point
		Conservative for multiple confounders

5 Real data examples

We applied the 5 QBA methods of Section 4 to data from the BCG and NHANES studies. In both examples the naive analysis was the linear regression $Y|X, C$ with binary exposure X. We used measured variables to represent the unmeasured confounders U. So, in effect our analyses examined the effect of not including certain confounders and we assumed that after adjustment for U and C there was

no unmeasured confounding. For the BCG example, U was a single confounder and adjustment for U did not change the study conclusions. In contrast for the NHANES example, U represented multiple confounders which did affect the study conclusions.

For *treatSens* we used Probit regression for its treatment model because X was binary, and for *causalsens* we used the one-sided confounding function because we assumed the exposure effect was the same in both exposure groups. Using measured covariates C, we calculated benchmark values the E-values of the point estimate and CI limit, and for the bias parameters, ϕ, of the other four programs.

As these are illustrative examples (of QBA to unmeasured confounding) we have ignored other potential sources of bias (such as missing data) and only considered a small number of measured covariates. We restricted our analyses to participants with complete data on Y, X, C and U.

We introduce each example, separately describe each QBA's results and then conclude with a summary across of 5 methods.

5.1 Example 1: BCG Study

The BCG study is a follow-up of a dietary intervention randomized controlled trial of pregnant women and their offspring [22,23]. Data were collected on the offspring (gestational age, sex, and 14 weight and height measures at birth, 6 weeks, $3, 6, 9$ and 12 months, and thereafter at 6-monthly intervals until aged 5 years) and their parents (anthropometric measures, health behaviours and socioeconomic characteristics). When aged 25, these offspring were invited to participate in a follow-up study in which standard anthropometric measures were recorded. We refer to the offspring, later young adults in the follow-up study, as the study participants.

Our analysis was a linear regression of adult body mass index (BMI) at age 25 on being overweight at age 5 years (BMI $\geq$ 17.44 kg/m^2 [52]). Measured covariates C were participant's sex and gestational age, and parents' height and weight measurements. We refer to maternal weight as the "strongest measured covariate" because it had the largest associations with child overweight and adult BMI. The unmeasured confounder U was a measure of childhood socioeconomic position (SEP) (paternal occupational social class based on the UK registrar general classification [53]). Based on the 544 participants with complete data on all variables, $\hat{\beta}_{X|C}$ was 2.28 kg/m^2 (95% confidence interval (CI) 1.39, 3.17 kg/m^2; P-value < 0.0001) and the fully adjusted estimate (i.e., adjusted for C and U) was 2.24 kg/m^2 (95% CI $1.35, 3.13$ kg/m^2; P-value < 0.0001). Statistical significance was defined at the 5% level.

5.1.1 Results from *treatSens*

The treatSens QBA results are shown in Figure 1(a). The axes represent values of the bias parameters $\phi = (\zeta^Z, \zeta^Y)$, where ζ^Z and ζ^Y denote the conditional associations between U and child overweight, and between U and adult BMI, respectively. Each contour represents the different combinations of ϕ that result in the same

bias-adjusted estimate, $\hat{\beta}_{X|C,U(\phi)}$. For example, $\hat{\beta}_{X|C,U(\phi)} = 0.18$ standard deviations of BMI when $\zeta^Y = 0.2$ and $\zeta^Z = 0.45$, and when $\zeta^Y = 0.4$ and $\zeta^Z = 0.25$. (Note that, *treatSens* standardises all continuous variables.) The black horizontal contour at $\zeta^Y = 0$ denotes the naive estimate of 0.51 standard deviations of BMI. The red contour represents the combinations of ϕ that would result in a null exposure estimate, and the blue contours bracket statistically insignificant exposure estimates. The pluses and inverted triangles denote the benchmark values of ϕ based on measured covariates C. The pluses denote confounders positively associated with adult BMI, and the inverted triangles denote confounders negatively associated with adult BMI, with those negative associations rescaled by -1. The red cross furthest away from the origin denotes the strongest measured covariate (maternal weight), and if U had a similar confounding effect to maternal weight then $\hat{\beta}_{X|C,U(\phi)}$ would be 0.49 standard deviations of BMI. The grey contour denotes the combinations of ϕ such that $\hat{\beta}_{X|C,U(\phi)} = 0.49$.

The contour plot suggests that a U similar to one of the measured covariates would at most reduce the point estimate by 3 or increase it by 4 standard deviations of BMI (i.e., $\hat{\beta}_{X|C,U(\phi)} \approx 0.55$ if ζ^Z was negative and of a similar magnitude to that of the association between child overweight and maternal weight). Furthermore, in order for U to change the study conclusions (such as explain away $\hat{\beta}_{X|C}$ or its statistical significance, or double the value of $\hat{\beta}_{X|C}$) then U would need to be a far stronger confounder than the strongest measured covariate (i.e., the magnitudes of ζ^Z and ζ^Y would need to be more than double the associations between maternal weight and child overweight and adult BMI, respectively).

5.1.2 *causalsens*

Figure 1(b) shows the results of the *causalsens* QBA where the amount of unmeasured confounding and its direction of effect is represented by the directional proportion R^2_α. The black line represents the bias-adjusted exposure estimates, the grey shaded area the corresponding 95% confidence intervals, and the crosses are the benchmark partial R^2 values based on measured covariates C (n.b., each benchmark is depicted as having a negative and positive direction of effect). Values of $R^2_\alpha > 0$ corresponds to individuals in the exposed group tending to have higher potential outcomes (to both exposure to being overweight and not overweight at age 5) than individuals in the unexposed group (i.e., unexposed group are healthier); and the converse for $R^2_\alpha < 0$.

If the residual variance explained by U was comparable to that of the strongest measured covariate (i.e., $|R^2_\alpha| \approx 5.5\%$) then $\hat{\beta}_{X|C,U(\phi)}$ could be as large as 5 kg/m^2 (with P-value < 0.05) or close to the null. For weaker levels of confounding ($|R^2_\alpha| < 0.01$), $\hat{\beta}_{X|C,U(\phi)}$ could be between 1.25 and 3.50 kg/m^2. To explain away statistical significance, the unexposed group would need to be healthier than the exposed group (regardless of exposure status) and U would need to account for at least 2% of the residual variance of adult BMI (i.e., $R^2_\alpha \geq 0.02$)).

5.1.3 *sensemakr*

Figures 1(c) and (d) show the contour plots of the QBA results for the point estimate and t-value, respectively, when U is assumed to reduce the point estimate. The axes represent values of the bias parameters $R^2_{X\sim U|C}$ (partial R^2 of U with exposure X) and $R^2_{Y\sim U|X,C}$ (partial R^2 of U with outcome Y). The contours have a similar interpretation as discussed for *treatSens*. For example, the red contour represents different combinations of $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$ that result in a null point estimate (Figure 1(c)) and t-value corresponding to 5% statistical significance (Figure 1(d)). The black triangle denotes the naive result and the red diamonds denote once, twice and thrice the benchmark values based on the strongest measured covariate (maternal weight). Results indicate that the proportion of residual variance of child overweight and adult BMI explained by U would need to be more than 3 times that of maternal weight in order for unmeasured confounding to either explain away all of $\hat{\beta}_{X|C}$ or its statistical significance. If the magnitude of the confounding effect of U was comparable to that of maternal weight, and U reduced the point estimate, then $\hat{\beta}_{X|C,U(\phi)}$ would be about 1.96 kg/m^2 with corresponding t-value of 4.44 (P-value < 0.0001).

The robustness values for $\hat{\beta}_{X|C}$ and its statistical significance were 19.47% and 12.36%, respectively. So, U could explain away all of $\hat{\beta}_{X|C}$ (or its statistical significance) if U accounted for at least 19.47% (or 12.36%) of the residual variance of both child overweight and adult BMI after conditioning on C. In keeping with the contour plots, these robustness values were substantially higher than the benchmark values for $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$ (Supplementary Table S1) indicating that the confounding effect of U would need to be far stronger than that of even the strongest measured covariate in order for unmeasured confounding to reduce the exposure effect to the null or remove its statistical significance.

Supplementary Figures S1(a) and (b) show the corresponding contour plots for the point estimate and t-value, respectively, when U is assumed to increase the point estimate. If the magnitude of the confounding effect of U was comparable to that of maternal weight, and U increased the point estimate, then $\hat{\beta}_{X|C,U(\phi)}$ would be about 2.60 kg/m^2 with corresponding t-value of 5.02 (P-value < 0.0001).

5.1.4 E-value

The E-value for a null exposure effect was 2.55 indicating that the exposure effect after adjusting for C and U could be null (or in the reverse direction) if associations between U and child overweight and between U and adult BMI, after conditioning on C, exceeded 2.55 on the risk ratio scale (Supplementary Figure S2). For slightly weaker associations with U, ≥ 1.98 but < 2.55, the exposure effect estimate would remain positive but be statistically insignificant. The benchmark E-values (Supplementary Table S2) for the point estimate and lower CI limit were comparable to the E-values for $\hat{\beta}_{X|C}$ and its CI limit. Omitting the strongest measured covariate resulted in a small increase of both E-values.

5.1.5 *konfound*

The percent bias and impact threshold were 60.1% (Supplementary Figure S3) and 0.139 (Supplementary Figure S4), respectively. Therefore, in order for unmeasured confounding to explain away the statistical significance of $\hat{\beta}_{X|C}$ then (1) U would need to account for at least 60.1% of $\hat{\beta}_{X|C}$ (i.e., giving an exposure estimate adjusted for C and U of $\leq$ 1.37 kg/m^2) and (2) the partial correlations of U with adult BMI and child overweight must both exceed 0.373 (i.e., $\sqrt{0.139}$). If the partial correlations of U with adult BMI and child overweight were comparable to those of the measured covariates (Supplementary Table S3) then we would expect the exposure effect to remain statistically significant even after adjusting for U.

5.1.6 Summary

The results from *treatSens*, *sensemakr*, *E-value* and *konfound* indicate that if U was comparable to even the strongest measured covariate then we would still conclude that children overweight at age 5 years tended to have a higher BMI in young adulthood. Furthermore, *treatSens* and *sensemakr* indicated that the strength of the $X - U$ and $Y - U$ associations would need to be at least double those of the strongest measured covariate in order for unmeasured confounding to substantially change the study conclusions (i.e., statistically insignificant effect, or a null or reversed effect). In contrast, from *causalsens* we conclude that unmeasured confounding could substantially change the study conclusions (i.e., if U was comparable to the strongest measured covariate then we cannot exclude the possibility that being overweight at age 5 years had no (or little effect) on a person's BMI in young adulthood).

5.2 Example 2: NHANES study

The NHANES study consists of a series of health and nutrition surveys conducted by the National Center for Health Statistics. Every year since 1999, approximately $5,000$ individuals of all ages are interviewed in their homes with health examinations conducted in a mobile examination centres. We analysed data from the $2015-2016$ NHANES survey.

Our analysis was a linear regression of systolic blood pressure (SBP) on diabetes among adults aged $\geq$ 18 years. Diabetes was defined as a HbA1c measurement of at least 6.5% (diabetes= 1 if HbA1C $\geq$ 6.5%, 0 otherwise) [54]. Measured covariates were age and sex, with age as the strongest measured covariate (i.e., largest associations with diabetes and SBP). The unmeasured confounders were BMI, ethnicity and poverty income ratio (PIR; the ratio of family income to the federal poverty line [55]). Based on the $4,576$ participants with complete data on all variables, $\hat{\beta}_{X|C}$ was 3.48 mmHg (99% CI $1.55, 5.40$ mmHg; P-value < 0.0001) and the fully adjusted estimate was 1.67 mmHg (99% CI $-0.27, 3.61$ mmHg; P-value 0.03). So, controlling for BMI, ethnicity and PIR explained 48% of $\hat{\beta}_{X|C}$ and resulted in a 99% CI that contained the null (i.e., P-value greater than 0.01). Statistical significance was defined at the 1% level.

Note that, we applied *sensemakr* with age and sex as a grouped benchmark, and had to edit the source code of *causalsens* in order to change the statistical significance level from 5% to 1%.

5.2.1 *treatSens*

Figure 2(a) shows the results of the *treatSens* QBA. If the magnitudes of the diabetes$-U$ and SBP-U associations were comparable to those of the strongest measured covariate, age (diabetes$-$age$=$ 0.45, and SBP$-$age$=$ 0.44 on the standardised scale) then $\hat{\beta}_{X|C,U(\phi)}$ could be ≈ 0.28 standard deviations of SBP (47% increase of $\hat{\beta}_{X|C}$ and statistically significant) or approximately 0.105 standard deviations of SBP (47% reduction of $\hat{\beta}_{X|C}$) with a P-value of 0.01. Therefore, a confounder comparable to age could explain away the statistical significance of $\hat{\beta}_{X|C}$. For unmeasured confounding to explain away all of $\hat{\beta}_{X|C}$ then U would need to have stronger associations with either diabetes, SBP or both (e.g., double that of diabetes$-$age ($\zeta^z \approx 1, \zeta^y \approx 0.44$), double that of SBP$-$age ($\zeta^z \approx 0.45, \zeta^y \approx 1$), or in-between for both ($\zeta^z \approx 0.55, \zeta^y \approx 0.75$)).

5.2.2 *causalsens*

Figure 2(b) shows the results of the *causalsens* QBA, where $R^2_\alpha > 0$ corresponds to individuals in the diabetic group tending to have higher potential SBP values (to both exposure to diabetes and no exposure) than the non-diabetic group (i.e., healthier individuals were non-diabetic); and the converse for $R^2_\alpha < 0$. Note that, the default scale for $R^2_\alpha > 0$ excluded the benchmark for the strongest measured confounder (age).

If the residual variance explained by U was comparable to that of the weakest measured covariate, sex, ($|R^2_\alpha| = 0.0108$) then $\hat{\beta}_{X|C,U(\phi)}$ could be as large as 10 mmHg or a reversed effect of about -2 mmHg; both with a P-value < 0.01. And, if U had a partial R^2 value closer to that of age then $\hat{\beta}_{X|C,U(\phi)}$ could be ≤ -10 mmHg or ≥ 15 mmHg. In order to explain away all of $\hat{\beta}_{X|C}$ or its statistical significance, then U would need to explain a smaller proportion of the residual variance than sex and individuals in the non-diabetic group would need to be healthier than those of the diabetic group (regardless of diabetes status).

5.2.3 *sensemakr*

The robustness values for $\hat{\beta}_{X|C}$ and 1% statistical significance were 6.65% and 4.36%, respectively. Since Supplementary Table S4 shows at least one of the measured covariates accounts for at least 6.65% of the residual variation of diabetes and SBP, then we cannot exclude the possibility that unmeasured confounding could explain away all of $\hat{\beta}_{X|C}$ or all of its 1% statistical significance. This is supported by Figures 2(c) and (d) which show that even if U was a weaker confounder than age, provided the direction of its effect was to reduce the point estimate, then accounting for U could result in a null or statistically insignificant exposure effect. Depending on the direction of the effect of U, if the magnitude of the confounding

effect of U was comparable to age then the exposure effect could be reversed with $\hat{\beta}_{X|C,U(\phi)} = -3.41$ mmHg (Figure 2(c)) or increased to $\hat{\beta}_{X|C,U(\phi)} = 15.73$ mmHg (Supplementary Figure S5(a)).

5.2.4 *E-value*

The E-value for a null exposure effect was 1.67, and the E-value for a P-value ≤ 0.01 was 1.38 (Supplementary Figure S6). These risk ratios are not implausibly large and so indicate that the exposure effect could be null or statistically insignificant after adjusting for U. The benchmark E-values for age were noticeably larger (Supplementary Table S5) than the E-values.

5.2.5 *konfound*

According to *konfound*, controlling for U could result in a statistically insignificant exposure effect (at the 1% level) if U explained away at least 44.67% of $\hat{\beta}_{X|C}$ (i.e., $\hat{\beta}_{X|C,U(\phi)} < 1.93$mmHg; Supplementary Figure S3) or the magnitude of the partial correlations of U with SBP and diabetes both exceeded 0.178 (Supplementary Figure S4). The benchmark partial correlations (Supplementary Table S7) for age were noticeably larger than 0.178.

5.2.6 Summary

If U were comparable to the strongest measured covariate age then *causalsens*, *sensemakr*, and *E-value* indicated that the exposure effect adjusted for C and U would either be null or in the reverse direction, while *treatSens* suggested that the exposure effect would still be positive although not statistically significant at the 1% level. *konfound* also indicated that the statistical significance of the exposure effect was not robust to unmeasured confounding. Given there were only two measured covariates, it seems plausible that unmeasured confounding would at least change the study conclusions with respect to statistical significance, and possibly also the direction of the exposure effect.

6 Discussion

We have conducted an up-to-date review of software implementations of QBA to unmeasured confounding, and a comparative evaluations of 5 different implementations applicable for a linear regression analysis. Our review reported many new QBA software programs since the last published review [10], most of which are implemented in the freely available statistical software environment R. Many programs include features such as benchmarking and graphical displays of the QBA results to aid interpretation. Our comparative evaluation illustrated the wide variation in the types of QBA methods applicable to a linear regression analysis. Methods that only indicated sensitivity at a tipping point (i.e., E-value and *konfound*) were less informative than those that provided bias-adjusted results across a range of scenarios (i.e, *treatSens*, *sensemakr* and *causalsens*).

When the unmeasured confounder (or confounders) are known the analyst can obtain information about the potential values of the bias parameters from external sources such as published studies. Obtaining this external information can be difficult for parameters that are not routinely reported; for example, when the analysis of interest is a linear regression, studies tend to report regression coefficients and not partial R^2 values. Therefore, it is generally easier to find external information on QBA methods such as *treatSens* (where the bias parameters are coefficients of a regression) compared to those of methods such as *sensemakr*, *causalsens*, and *konfound*.

Benchmarking is a useful tool to aid researchers in judging the plausibility of the values of the bias parameters, especially when the identity of the unmeasured confounders is unknown, or external information is unattainable. Generally, benchmarking assumes that U has similar confounding properties to a measured covariate C_j (or group of covariates). Ideally, these benchmark values should be estimated after adjustment for the omission of the unmeasured confounder(s) [24,38]. However, several QBA methods (such as *treatSens*, *causalsens* and *konfound* and others [56, 57]) calculate their benchmarks based on the naive model $Y|X, C$; for example, using the coefficient for covariate C_j from the naive model $Y|X, C$ (for $C = (C_j, C_{-j})$ as a benchmark for the conditional association of $Y - U$ given X and C. Omitting U when estimating a benchmark can change its value even when C and U are independent, which can lead to incorrect conclusions about the sensitivity of the exposure effect to unmeasured confounding [24, 38]. Benchmark methods that adjust for C have been proposed, including *sensemakr* which defines upper bounds for its benchmarks [24, 38, 58]. Benchmarking cannot be applied in a meaningful way to QBA methods based on an extreme scenario (such as the *E-value*) [24].

Examples of QBAs tend to focus on a single unmeasured confounder when in fact many weaker unmeasured confounders can jointly change a study's conclusions [4]. However, several QBA methods are generalisable to multiple unmeasured confounders without burdening the analyst with additional bias parameters. For example, a common assumption is that U represents a linear combination of multiple unmeasured confounders, with the elementary scenario that U is a single unmeasured confounder. A drawback of this appealing assumption is that the QBA tends to be conservative for multiple unmeasured confounders [38]. Alternatively, a QBA method may leave the functional form of U unspecified and instead define its bias parameters as upper bounds (such as the *E-value* where U is a categorical variable with categories representing all possible combinations of the multiple unmeasured confounders and its bias parameters RR_{XU} and RR_{UY} are the maximum risk ratios comparing any two categories of U [59]). A drawback of these upper bounds is that they correspond to extreme situations, making it hard to locate appropriate benchmarks values or external information. To address both drawbacks, a QBA could explicitly model each unmeasured confounder separately whilst allowing for correlations between the confounders, although this would then increase the number of bias parameters. If many unmeasured confounders are suspected, then the analyst should question if a QBA is suitable since the accuracy of a QBA generally relies on a study having measured key confounders. Importantly, a QBA is not a

replacement for a correctly designed and conducted study.

In our review, all software implementations were of deterministic QBA methods. In general, deterministic QBA are tipping point analyses with statistical significance as one of the tipping points. Given the call to move away from reliance on statistical significance [60], we recommend QBA methods that provide bias-adjusted results for all specified values of the bias parameters to give a complete picture of the effect of unmeasured confounding (such as *treatSens*, *sensemakr* and *causalsens*). However, presenting and interpreting these results can be challenging, especially when there are more than two bias parameters due to the large number of possible value combinations (e.g., three parameters each with 10 possible values gives 1000 combinations). An alternative is a probabilistic QBA which summarises the results as a point estimate and accompanying interval estimate. The advantages of the probabilistic QBA are: (1) the output is familiar to epidemiologists (i.e., similar to point estimate and 95% CI), (2) the interval estimate accounts for all sources of uncertainty due to bias and random sampling, and (3) less reliance on the statistical significance interpretation. Further work is needed to provide software implementations of probabilistic QBAs.

In summary, there have been several new software implementations of QBAs, most of which are available in R. And our comparative evaluation has illustrated the wide diversity in the types of QBA method that can be applied to the same substantive analysis of interest. Such diversity of QBA methods presents challenges in the widespread uptake of QBA methods. Guidelines are needed on the appropriate choice of QBA method, along with provision of software implementations in platforms other than R.

Acknowledgements

We thank the study executives of NHANES, and Dr. P. C. Elwood (MRC Epidemiology Unit, South Wales) and Prof. Y. Ben-Shlomo (University of Bristol) for permitting access to the BCG study data. RAH and EK are supported by a Sir Henry Dale Fellowship that is jointly funded by the Wellcome Trust and the Royal Society (grant 215408/Z/19/Z), and KT works in the MRC Integrative Epidemiology Unit, which is supported by the University of Bristol and the Medical Research Council (grants MC_UU_00011/3).

References

[1] Hernán M, Robins J. *Causal inference: What if.* 2020.

[2] Arah OA. Bias analysis for uncontrolled confounding in the health sciences. *Annual review of public health.* 2017;38:23–38.

[3] Fewell Z, Davey Smith G, C SJA. The Impact of Residual and Unmeasured Confounding in Epidemiological Studies: a Simulation Study. *Americal Journal of Epidemiology.* 2007;166(6):646–655.

[4] Groenwold RH, Sterne JA, Lawlor DA, Moons KG, Hoes AW, Tilling K. Sensitivity analysis for the effects of multiple unmeasured confounders. *Annals of epidemiology*. 2016;26(9):605–611.

[5] Uddin MJ, Groenwold RH, Ali MS, de Boer A, Roes KC, Chowdhury MA, Klungel OH. Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *International journal of clinical pharmacy*. 2016; 38(3):714–723.

[6] Lash TL, Fox MP, Fink AK. *Applying quantitative bias analysis to epidemiologic data*. Springer Science & Business Media. 2009.

[7] Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *International journal of epidemiology*. 2014;43(6):1969–1985.

[8] Hunnicutt JN, Ulbricht CM, Chrysanthopoulou SA, Lapane KL. Probabilistic bias analysis in pharmacoepidemiology and comparative effectiveness research: a systematic review. *Pharmacoepidemiology and drug safety*. 2016;25(12):1343–1353.

[9] Liu W, Kuramoto SJ, Stuart EA. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science*. 2013;14(6):570–580.

[10] Peel MJ. Addressing unobserved endogeneity bias in accounting studies: control and sensitivity methods by variable type. *Accounting and Business Research*. 2014;44(5):545–571.

[11] Streeter AJ, Lin NX, Crathorne L, Haasova M, Hyde C, Melzer D, Henley WE. Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *Journal of clinical epidemiology*. 2017;87:23–34.

[12] Zhang X, Faries DE, Li H, Stamey JD, Imbens GW. Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiology and drug safety*. 2018;27(4):373–382.

[13] Arah OA, Chiba Y, Greenland S. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. *Annals of Epidemiology*. 2008; 18:637–646.

[14] Groenwold RH, Nelson DB, Nichol KL, Hoes AW, Hak E. Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. *International journal of epidemiology*. 2010;39(1):107–117.

[15] MacLehose RF, Ahern TP, Lash TL, Poole C, Greenland S. The importance of making assumptions in bias analysis. *Epidemiology (Cambridge, Mass)*. 2021; 32(5):617.

[16] McCandless LC, Gustafson P. A comparison of Bayesian and Monte Carlo sensitivity analysis for unmeasured confounding. *Statistics in medicine*. 2017; 36(18):2887–2901.

[17] Mittinty MN. Estimation bias due to unmeasured confounding in oral health epidemiology. *Community Dental Health*. 2020;37:1–6.

[18] Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and drug safety*. 2006;15(5):291–303.

[19] Steenland K, Greenland S. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *American Journal of Epidemiology*. 2004;160(4):384–392.

[20] Thommes EW, Mahmud SM, Young-Xu Y, Snider JT, van Aalst R, Lee JK, Halchenko Y, Russo E, Chit A. Assessing the prior event rate ratio method via probabilistic bias analysis on a Bayesian network. *Statistics in medicine*. 2020;39(5):639–659.

[21] for Disease Control C, for Health Statistics (NCHS) PCNC. National Health and Nutrition Examination Survey (NHANES) 2015-2016. *https://wwwncdcgov/nchs/nhanes/continuousnhanes/defaultaspx?BeginYear=2015*. 2016;.

[22] Elwood PC, Haley T, Hughes S, Sweetnam P, Gray O, Davies D. Child growth (0-5 years), and the effect of entitlement to a milk supplement. *Archives of Disease in Childhood*. 1981;56(11):831–835.

[23] McCarthy A, Hughes R, Tilling K, Davies D, Davey Smith G, Ben-Shlomo Y. Birth weight; postnatal, infant, and childhood growth; and obesity in young adulthood: evidence from the Barry Caerphilly Growth Study. *The American journal of clinical nutrition*. 2007;86(4):907–913.

[24] Zhang B, Small DS. A calibrated sensitivity analysis for matched observational studies with application to the effect of second-hand smoke exposure on blood lead levels in children. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2020;69(5):1285–1305.

[25] Greenland S. The sensitivity of a sensitivity analysis. In: *Proceedings-American Statistical Association Biometrics Section*. UNKNOWN. 1997; pp. 19–21.

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 2021.
URL `https://www.R-project.org/`

[27] StataCorp. Stata Statistical Software: Release 17. *College Station, TX StataCorp LLC*. 2021;.

[28] Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for R, 2020. *URL https://CRAN R-project org/package= shiny R package version*. 2021;1(0):p517.

[29] Gorbach T, de Luna X. Inference for partial correlation when data are missing not at random. *Statistics & Probability Letters*. 2018;141:82–89.

[30] Blackwell M. A selection bias approach to sensitivity analysis for causal effects. *Political Analysis*. 2014;22(2):169–182.

[31] VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Annals of internal medicine*. 2017;167(4):268–274.

[32] Mathur MB, Ding P, Riddell CA, VanderWeele TJ. Website and R package for computing E-values. *Epidemiology (Cambridge, Mass)*. 2018;29(5):e45.

[33] Linden A, Mathur M, VanderWeele T. *Evalue: Stata module for conducting sensitivity analyses for unmeasured confounding in observational studies. Statistical software components S458592*. Department of Economics, Boston College. 2019.

[34] Mathur MB, VanderWeele TJ. Sensitivity analysis for unmeasured confounding in meta-analyses. *Journal of the American Statistical Association*. 2020; 115(529):163–172.

[35] Harada M. Generalized sensitivity analysis and application to quasi-experiments. *Tech. rep.*, Working Paper, New York University. 2013.

[36] Harada M. ISA: Stata module to perform Imbens'(2003) sensitivity analysis. 2012;.

[37] Xu R, Frank KA, Maroulis SJ, Rosenberg JM. konfound: Command to quantify robustness of causal inferences. *The Stata Journal*. 2019;19(3):523–550.

[38] Cinelli C, Hazlett C. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2020;82(1):39–67.

[39] Cinelli C, Ferwerda J, Hazlett C. sensemakr: Sensitivity Analysis Tools for OLS in R and Stata. *Submitted to the Journal of Statistical Software*. 2020;.

[40] Small DS, Cheng J, Halloran ME, Rosenbaum PR. Case definition and design sensitivity. *Journal of the American Statistical Association*. 2013; 108(504):1457–1468.

[41] Rosenbaum PR. *sensitivitymv: Sensitivity Analysis in Observational Studies*. 2018. R package version 1.4.3.
URL `https://CRAN.R-project.org/package=sensitivitymv`

[42] Rosenbaum PR. *sensitivitymw: Sensitivity analysis using weighted M-statistics*. 2014. R package version 1.1.
URL `https://CRAN.R-project.org/package=sensitivitymw`

[43] Lee K, Small DS, Rosenbaum PR. A powerful approach to the study of moderate effect modification in observational studies. *Biometrics*. 2018;74(4):1161–1170.

[44] Carnegie NB, Harada M, Hill JL. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*. 2016;9(3):395–420.

[45] Dorie V, Harada M, Carnegie NB, Hill J. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*. 2016;35(20):3453–3470.

[46] Franks A, D'Amour A, Feller A. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*. 2019;pp. 1–33.

[47] Genbäck M, de Luna X. Causal inference accounting for unobserved confounding after outcome regression and doubly robust estimation. *Biometrics*. 2019; 75(2):506–515.

[48] Cinelli C, Kumor D, Chen B, Pearl J, Bareinboim E. Sensitivity analysis of linear structural causal models. In: *International Conference on Machine Learning*. 2019; pp. 1252–1261.

[49] Robins JM. Association, causation and marginal structural models. *Synthese*. 1999;121:151–179.

[50] Robins JM. *Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models, Section 6–11. In Statistical Models in Epidemiology: The Environment and Clinical Trials*. Springer–Verlag: New York. 1999.

[51] Rubin DB. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*. 2005;100(469):322–311.

[52] Cole TJ, Bellizzi MC, Flegal KM, Dietz WH. Establishing a standard definition for child overweight and obesity worldwide: international survey. *Bmj*. 2000; 320(7244):1240.

[53] Pevalin D, Rose D. The national statistics socio-economic classification: unifying official and sociological approaches to the conceptualisation and measurement of social class in the United Kingdom. *Sociétés contemporaines*. 2002; (1):75–106.

[54] Geneva WHO. Use of glycated haemoglobic (HbA1c) in the diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. *Tech. rep.*, World Health Organization. 2011.
URL `https://www.ncbi.nlm.nih.gov/books/NBK304267/`

[55] Alaimo K, Briefel RR, Frongillo EA, Olson CM. Food insufficiency exists in the United States: results from the Third National Health and Nutrition Examination Survey. *American Journal of Public Health*. 1998;88(3):419–426.

[56] Imbens GW. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*. 2003;93(2):126–132.

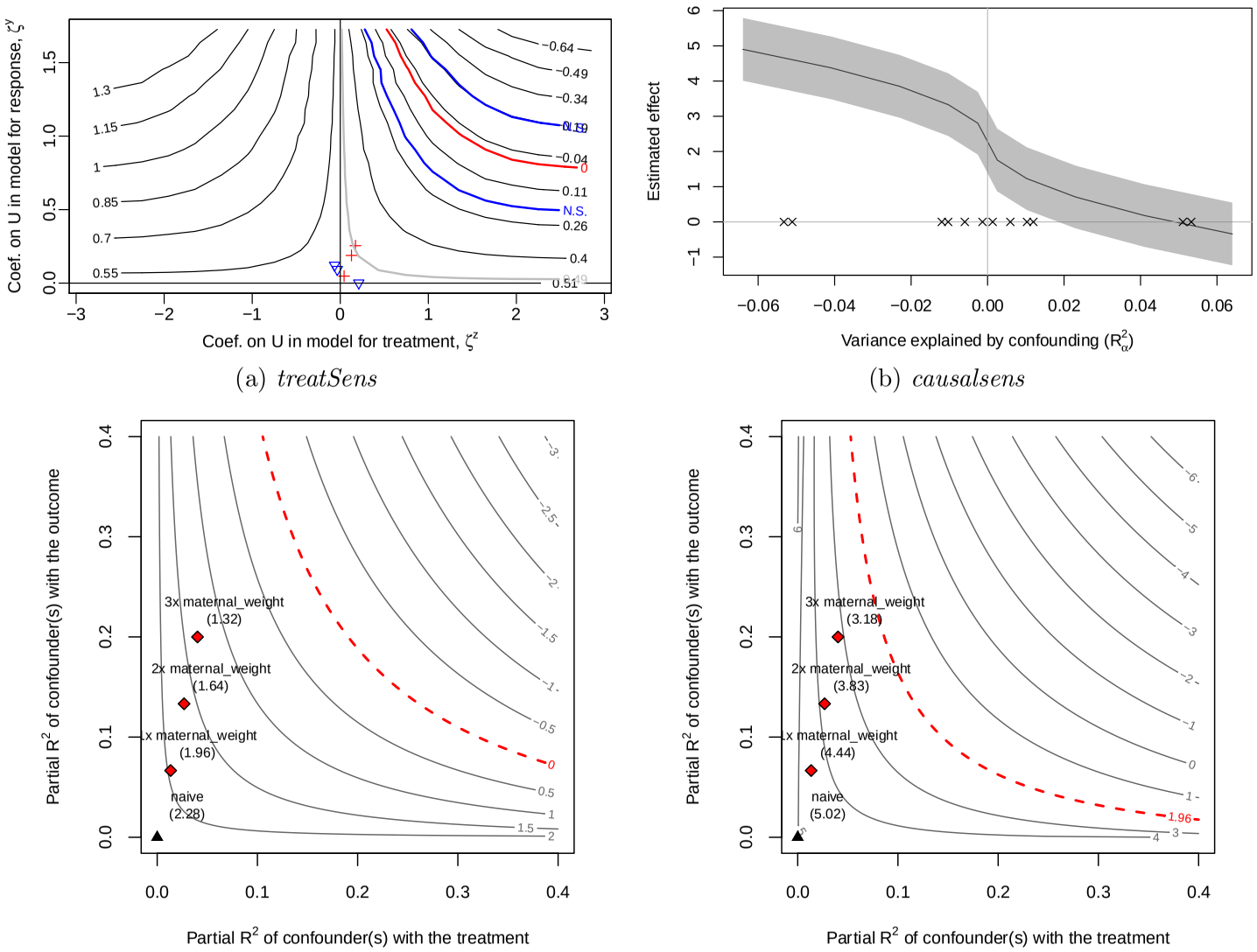
[57] Middleton JA, Scott MA, Diakow R, Hill JL. Bias amplification and bias unmasking. *Political Analysis*. 2016;24(3):307–323.

[58] Hsu JY, Small DS. Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics*. 2013;69(4):803–811.

[59] Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass)*. 2016;27(3):368.

[60] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567:305–307.

Figure 1: Quantitative bias analysis for effect of child overweight on adult body mass index from the Barry Caerphilly Growth study. Red contour (null effect in (a) and (c), t-value at 5% significance in (d)), blue contours (bracket 5% statistically insignificant estimates), black contour or line (bias-adjusted estimates), grey shaded area (95% confidence intervals for bias-adjusted estimates), pluses, inverted triangles, crosses, and diamonds (benchmarks), and black triangle (naive estimate).



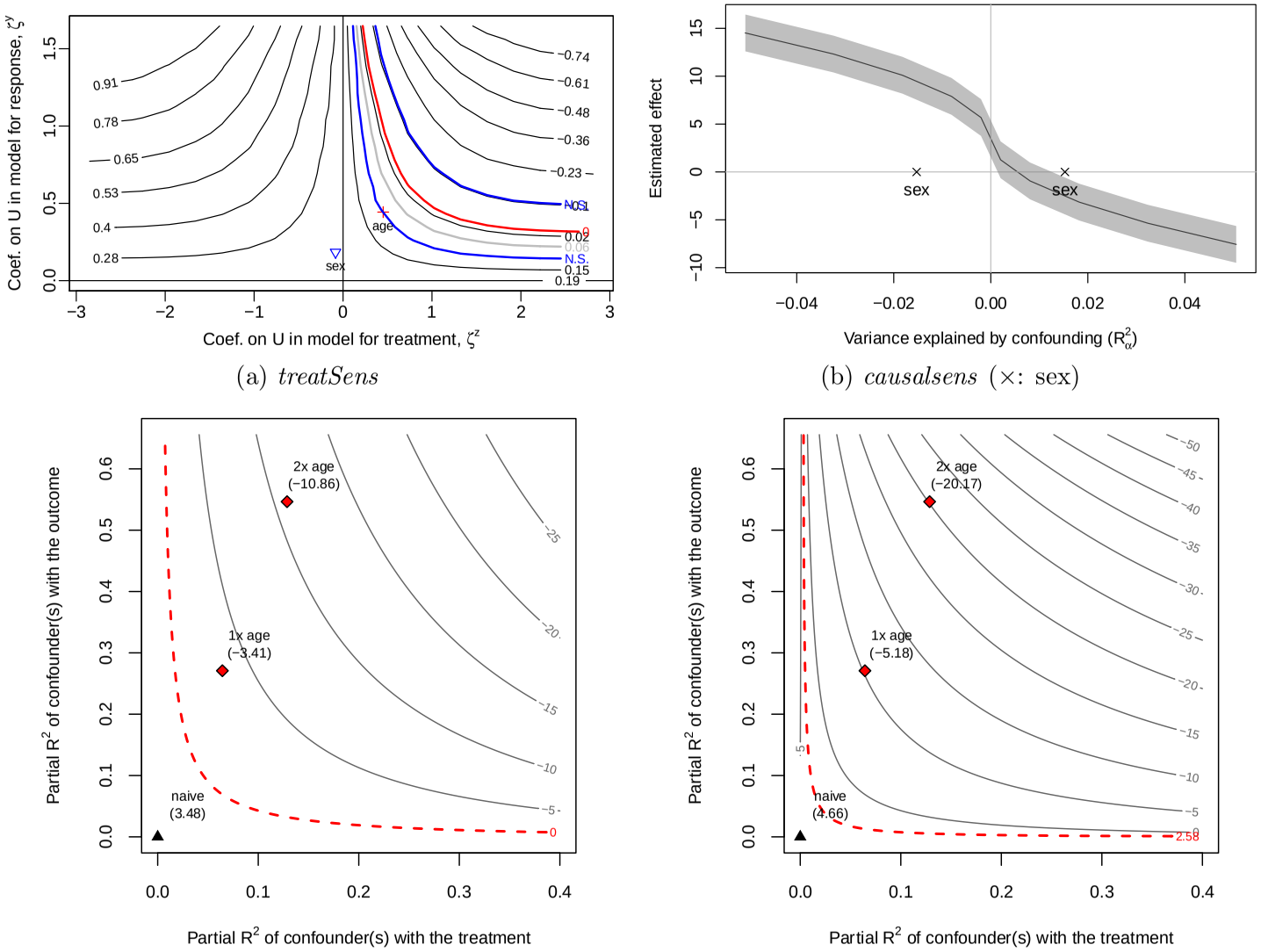
(a) *treatSens*

(b) *causalsens*

(c) *sensemakr* for point estimate

(d) *sensemakr* for t-value

Figure 2: Quantitative bias analysis for effect of diabetes on systolic blood pressure from the National Health and Nutrition Examination Survey. Red contour (null effect in (a) and (c), t-value at 1% significance in (d)), blue contours (bracket 1% statistically insignificant estimates), black contour or line (bias-adjusted estimates), grey shaded area (99% confidence intervals for bias-adjusted estimates), pluses, inverted triangles, crosses, and diamonds (benchmark), and black triangle (naive estimate).



(a) *treatSens*

(b) *causalsens* (×: sex)

(c) *sensemakr* for point estimate

(d) *sensemakr* for t-value