Journal Title
XX(X):2–27

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

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]

Abstract

Failure to appropriately account for unmeasured confounding may lead to erroneous conclusions. Quantitative bias analysis (QBA) can be used to quantify the potential impact of unmeasured confounding or how much unmeasured confounding would be needed to change a study's conclusions. Currently, QBA methods are not routinely implemented, partly due to a lack of knowledge about accessible software. We review the latest developments in QBA software between 2011 to 2021 and compare five different programs applicable when fitting a linear regression: *treatSens*, *causalsens*, *sensemakr*, *EValue*, and *konfound*. We illustrate application of these programs to two datasets and provide code to assist analysts in future use of these software programs. Our review found 21 programs with most created post 2016. All are implementations of a deterministic QBA, and the majority are available in the free statistical software environment R. Many programs include features such as benchmarking and graphical displays of the QBA results to aid interpretation. Out of the five programs we compared, *sensemakr* performs the most detailed QBA and includes a benchmarking feature for multiple unmeasured confounders. The diversity of QBA methods presents challenges to the widespread uptake of QBA among applied researchers. Provision of detailed QBA guidelines would be beneficial.

Keywords

Causal inference; Linear regression; Review; Sensitivity analysis; Software; Unmeasured confounding

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

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 the adjustment model is correct and a sufficient set of confounders has been measured without error[1]. Failure to appropriately account for unmeasured or poorly measured confounders in analyses may lead to invalid inference[2–4].

There are several approaches to assess causality which depend on assumptions other than "no unmeasured confounding" (e.g., self-controlled study designs, prior event rate ratio, instrumental variable analysis, negative controls, perturbation variable analysis, and methods that use confounder data collected on a study sub-sample[5]). When none of these approaches are applicable (e.g., study lacks an appropriate instrument or sub-sample data on the unmeasured confounders) then the analyst must assess the sensitivity of the study's conclusions to the assumption of no unmeasured confounding using a quantitative bias analysis (QBA; also known as a sensitivity analysis). A QBA can be used to quantify the potential impact of unmeasured confounding on an exposure effect estimate or to quantify how much unmeasured confounding would be needed to change a study's conclusions.

Currently, QBA methods are not routinely implemented. A recent published in 2016 found that the use of QBA for unmeasured confounding had not increased in the years $2010-2012$ compared to the $2004-2007$ period[6]. Lack of knowledge about QBA, and of analyst-friendly methods and software have been identified as barriers to the widespread implementation of a QBA[7–9]. In the past decade, there have been several reviews of QBA methods[2,5,9–17]. Only two of these papers reviewed software implementations of QBA methods: the supplementary of[15] provided a brief summary of software implementing Rosenbaum-style QBA methods[18], and[11] reviewed software implementations before its publication in July 2014. Also, comparisons of QBA methods have primarily been limited to analyses with a binary outcome[10,19–26].

Our paper reviews available software implementing a QBA to address unmeasured confounding caused by a study not collecting data on these confounders as opposed to mismeasurement of measured confounders. We then describe, illustrate and compare QBA software applicable when the analysis of interest is a linear regression. We illustrate how to apply these methods using a real-data example from the Barry

[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

Corresponding author:
Emily Kawabata, Oakfield House, Oakfield Grove, Bristol, BS8 2BN, United Kingdom
Email: emily.kawabata@bristol.ac.uk

Prepared using sagej.cls

Caerphilly Growth (BCG) study[27,28], and, in the Supplementary Materials, we also provide code implementing these methods when applied to publicly-accessible data from the $2015-2016$ National Health and Nutrition Examination Survey (NHANES) study[29].

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 (assuming no other sources of bias). 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 estimated from the observed data. Therefore, values for these bias parameters must be prespecified before conducting the QBA. Typically, the bias parameters specify the strength of the association between U and X given C, and between U and Y given X and C[21]. 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)[8], and from benchmarking (also known as calibration) where strengths of associations of measured covariates C with X and Y are used as benchmarks[30] for the bias parameters. 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 (CI) 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[7]. 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 prespecified 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 uses a prior probability distribution for ϕ to explicitly model the analyst's assumptions about which combinations of ϕ are most likely to occur and to incorporate their uncertainty about ϕ[7,22]. Averaging over this probability distribution generates a distribution of estimates of $\hat{\beta}_{X|C,U(\phi)}$ 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 prespecified probability) which accounts for uncertainty due to the unmeasured confounding and sampling variability[7].

3 Overview of available software

The aim of the literature review was to give a brief overview of publicly available software implementations of QBA, described in articles published between 1st January 2011 and 31st December 2021 (inclusive). We defined "software" to be either a web tool with a user-interface or software code that (i) was not specific to a particular data example (i.e., we excluded examples of code from empirical analyses), (ii) was freely available to download, and (iii) was accompanied by documentation detailing the software's features.

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 data extraction with any disagreements resolved by consensus. Eligible abstracts were published articles that either introduced a new QBA method or software implementation, compared or reviewed existing QBA methodology, or gave a tutorial on QBA. Examples of ineligible abstracts were meeting abstracts, commentaries, articles where a QBA was not conducted but mentioned as further work, and articles solely focused on answering applied questions (and so included limited information on the statistical methodology used). In stage 3, we extracted from the full text information about the analysis of interest, the QBA method, and the software used to implement the QBA.

After excluding duplicates, our Web of Science search identified 780 papers (flowchart of the review shown in Supplementary Figure S1). We excluded 24 meeting abstracts and editorials, 379 articles that did not conduct a QBA to unmeasured confounding, and 239 articles on applied analyses. Of the remaining 138, 29 articles referred to 21 publicly available software implementations of a QBA.

Table 1 summarises the key features of the 21 software programs in ascending date-order of creation. All 21 programs implement a deterministic QBA, with only 8 programs publicly available before 2017, and 17 implemented in the free software environment R[31]. Seven programs implement a QBA applicable for a matched observational study, five for a mediation analysis, and nine for a standard regression analysis. Five of the seven programs for a matched analysis (*sensitivityCaseControl, sensitivitymw, sensitivitymv, sensitivityfull* and *submax*) implement the same QBA method[18,32] but for different types of matched observational studies. For example, *sensitivitymw* is applicable to matched sets with one exposed subject and a fixed number of unexposed subjects, and *sensitivitymv* to matched sets with one exposed subject and a variable number of unexposed subjects. Also, *submax* and *sensitivityCaseControl* exploit effect modification and different definitions of a case of disease, respectively, to further evaluate sensitivity to unmeasured confounding. Among the programs for mediation analysis, *MediationSensitivityAnalysis* evaluates sensitivity to unmeasured confounding of the mediator-outcome relationship only, while the remaining programs can also evaluate sensitivity to unmeasured confounding of the exposure-mediator and exposure-outcome relationships.

Table 1. Software programs implementing a quantitative bias analysis for unmeasured confounding reported in articles published between 2011 and 2021.

		Analysis of interest				Bias analysis			
Name (Year created)	Environ-ment	Type of analysis	Outcome	Exposure	Mediator	No. bias parameters	Bench-marking	Graphical plot	Tipping point
isa[33,34] (2011)	Stata	simple[a]	con[b]	bin[c]	–	2	yes	line	null effect, stat. sig.[d]
gsa[11,35] (2012)	Stata	simple	bin, con	bin, con, cat[e]	–	2	yes	scatter	null effect, stat. sig.
SensitivityCase- Control[36,37] (2012)	R	matched	bin	bin	–	1	no	no	none
causalsens[38,39] (2013)	R	simple	con	bin	–	1	yes	line with CI[f]	null effect, stat. sig.
mbsens[40] (2014; 2)	Stata	matched	bin	bin	–	1	no	no	none
sensitivitymw[41–43] (2014)	R	matched	con, integer	bin	–	1	no	no	none
treatSens[44–46] (2014)	R	simple	con	bin, con	–	2 [+1 for $p(U)$][g]	yes	contour	null effect, stat. sig.
sensitivitymv[41,42,47] (2015)	R	matched	con, integer	bin	–	1	no	no	none
EValue[h](2017)[48–51]	R, Stata,	simple,	bin, con,	bin, con	–	2	no	line	null effect,

Continued on next page

Table 1 – continued from previous page

Name (Year created)	Environment	Analysis of interest: Type of analysis	Outcome	Exposure	Mediator	Bias analysis: No. bias parameters	Benchmarking	Graphical plot	Tipping point
	Web tool	meta-analysis	TTE[i]						stat. sig.
rmpw[52,53] (2017)	R	mediation	bin, con, cat	bin	bin, con, cat	2 or 4	yes	contour	stat. sig.
sensitivityfull[41,54,55] (2017)	R	matched	con, integer	bin	–	1	no	no	no
submax[56] (2017)	R	matched	con, integer	bin	–	1	no	no	stat. sig.
Umediation[57] (2017)	R	mediation	bin, con	bin, con	bin, con	3 [+1 or 2 for $p(U)$]	no	line	stat. sig.
konfound[58,59] (2018)	R, Stata, Web tool	simple	bin, con	bin, con	–	2	yes	bar, causal diagram	stat. sig.
sensmediation[60,61] (2018)	R	mediation	bin, con	bin, con	bin, con	1	no	line with CI	null effect, stat. sig.
sensitivityCalibration[30,62] (2018)	R	matched	con	bin	–	3	yes	line	stat. sig.
sensemakr[63,64] (2019)	R, Stata, Web tool	simple	con	bin	–	2	yes	contour	null effect, stat. sig.
ui[65] (2019)	R	simple	bin	bin	–	2	no	line with	null effect,

Continued on next page

Table 1 – continued from previous page

		Analysis of interest				Bias analysis			
Name (Year created)	Environ-ment	Type of analysis	Outcome	Exposure	Mediator	No. bias parameters	Bench-marking	Graphical plot	Tipping point
								CI and UI	stat. sig.
mediationsens[17,66] (2020)	R	mediation,	con, bin	bin	con, bin	2 or 3 [+1 for $p(U)$]	yes	contour	null effect, stat. sig.
survsens[67,68] (2020)	R	simple	TTE	bin	–	2 [+1 for $p(U)$]	no	contour	null effect, stat. sig.
MediationSensitivity-Analysis[69] (2021)	Web tool	mediation,	con	bin, con, cat	con	2	no	contour	null effect, stat. sig.

[a] estimation of total effect of exposure in a sample of unmatched, independent observations from a single study; [b]continuous variable; [c]binary variable; [d] statistical significance; [e]categorical variable; [f] CI: confidence interval; [g] option to set the parameter of the marginal distribution of U; [h]R package EValue, Stata command evalue and web tool E-value calculator; [i]time to event variable.

Most programs require the outcome (of the analysis of interest) to be either binary or continuous. However, program *survsens* implements a QBA specifically for a Cox proportional hazards regression analysis and is applicable for survival outcomes with or without competing risks. All programs can be applied to a binary exposure and seven programs are also applicable to a continuous or categorical exposure. Also, all programs allow the analysis of interest to adjust for measured covariates C of any variable type and generally assume that U represents the part of the unmeasured confounder(s) that is independent of C. Nine programs use the measured covariates to calculate benchmark values for the bias parameters.

The bias parameters represent the strength of the relationships between U and the exposure, outcome, or mediator. Programs *treatSens*, *Umediation*, *mediationsens*, and *survsens* also allow the analyst to vary the parameters of the marginal distribution of U (e.g., for binary U the probability $\Pr(U = 1)$). Otherwise, these marginal parameters are set to a default value (e.g., $\Pr(U = 1) = 0.5$).

Almost all programs report the values of the bias parameters at prespecified tipping points. Also, most programs output the bias-adjusted results (e.g., point estimate, CI or P-value for the exposure effect) at prespecified values of the bias parameter(s) (exceptions include *isa*, *gsa*, *konfound*, and R and Stata implementations of *EValue*). Note that, programs *uMediation* and *ui* summarise the bias-adjusted results using uncertainty intervals, which incorporates uncertainty about the values of the bias parameters and sampling variability. Fifteen programs generate a graphical plot of their QBA results.

Two programs also implement a QBA to other sources of bias: *MediationSensitivty-Analysis* can assess sensitivity to measurement error of the mediatior, outcome and measured covariates, and *Evalue* can assess sensitivity to differential misclassification of an outcome or exposure and to sample selection bias. Furthermore, both programs can simultaneously assess sensitivity to multiple sources of bias.

4. Quantitative bias analysis methods for linear regression

We describe and illustrate the following programs from Table 1 applicable for an unmatched analysis, where the exposure is binary and the exposure effect is estimated by a linear regression model: *treatSens*[44,45], *causalsens*[38], *sensemakr*[70], *EValue*[48], and *konfound*[58]. For reasons of brevity, we excluded programs *isa* and *gsa* as they are similar to the more recently published *treatSens*.

All five programs are implemented as an R package[39,46,49,59,64,71,72], and *sensemakr*[64], *EValue*[49,50] and *konfound*[58,73] are also available as a Stata command and web tool. Individual participant data is required for *treatSens* and *causalsens*, *EValue* only requires summary data from the naive analysis, and *sensemakr* and *konfound* can be applied to individual participant and summary data. Additionally, *treatSens*, *causalsens*, and *sensemakr* require prespecified values for ϕ which can be set by the analyst or set using the program's default values. Note that all five methods can be applied when $\hat{\beta}_{X|C}$ is not null, irrespective of whether $\hat{\beta}_{X|C}$ is statistically significant or not, and when $\hat{\beta}_{X|C}$ is null. However, for *treatSens* the tipping point for the point estimate is fixed at the null, and so this feature can only be used when $\hat{\beta}_{X|C}$ is not null.

Below is a summary of the five programs with further details in the Supplementary Materials.

4.1 treatSens

Program *treatSens* implements a simulation-based QBA[44] which is similar to multiple imputation for missing data[74]. For a prespecified value of ϕ, *treatSens* simulates U multiple times from the conditional distribution $U|Y, X, C$ given by the bias model. For each set of simulated values of U, the exposure effect is estimated from a linear regression of Y given X, C and the simulated U, and then Rubin's rules[74] are used to combine the multiple sets of results into a single estimate for $\hat{\beta}_{X|C,U(\phi)}$ and its standard error.

The bias model consists of three sub-models: the analysis model (i.e., linear regression of Y given X, C and U), the treatment model (e.g., linear or probit regression regression of X on C and U for continuous or binary X, respectively), and a marginal model for U (standard normal or Bernoulli distribution for continuous or binary U, respectively). It has two bias parameters $\phi = (\zeta^Y, \zeta^Z)$: ζ^Y is the coefficient for U from the analysis model $Y|X, C, U$ and ζ^Z is the coefficient of U from the treatment model $X|C, U$. To allow for bias in both directions (i.e., increased exposure effect, and reduced or reversed exposure effect), positive and negative values are specified for ζ^Z. The remaining coefficients of the treatment and analysis models are estimated from the observed data. The coefficients of measured covariates C (from the regressions of Y on X and C, and X on C) are used as benchmark values for ζ^Y and ζ^Z, respectively[44]. All continuous variables are standardised to facilitate comparison between these benchmark values and the bias parameters.

Program *treatSens* outputs a contour plot of the bias-adjusted estimates, $\hat{\beta}_{X|C,U(\phi)}$, for different combinations of ζ^Y and ζ^Z, indicating the values of ζ^Y and ζ^Z that correspond to tipping points for the point estimate (fixed at the null) and statistical significance (analyst can set the significance level; default is 5%). Additional outputs include tables of: (1) combination values of ζ^Y and ζ^Z at the tipping points, (2) $\hat{\beta}_{X|C,U(\phi)}$ and corresponding standard errors for prespecified values of ζ^Y and ζ^Z and each set of simulated U, and (3) benchmark values for ζ^Y and ζ^Z.

4.2 causalsens

Program *causalsens* generates a modified outcome, Y_ϕ^{adj}, which is adjusted for the bias due to unmeasured confounding for a prespecified value of ϕ[38]. The naive analysis is then refitted using Y_ϕ^{adj} instead of Y and the resulting exposure effect estimate and CI are the bias-adjusted results.

The QBA of *causalsens* is based on the potential outcomes framework[75]. Program *causalsens* requires a binary X, and so there are two potential outcomes per subject: $Y(0)$ when not exposed and $Y(1)$ when exposed. The bias model consists of a treatment model and a "confounding function"[76,77]. The treatment model is a logistic regression used to estimate the probability of being in the exposed group given C. The confounding function quantifies the average difference in potential outcomes $Y(0)$ (or $Y(1)$) between those in the exposed and unexposed groups, with any nonzero difference attributed to unmeasured confounding. It is parameterised by a single

bias parameter, $\phi = (R^2_\alpha)$: R^2_α denotes the proportion of unexplained variance in the potential outcomes that is explained by U and is set to positive and negative values to allow U to move the point estimate towards and away from the null. Program *causalsens* supplies two choices for the confounding function, named the "one-sided function" and the "alignment function", and also allows the analyst to specify their own function. The one-sided function assumes the true exposure effect is identical in the exposed and unexposed groups. Setting $R^2_\alpha > 0$ implies that the mean of $Y(1)$ (and $Y(0)$) is higher for the exposed group than the unexposed group, leading $\hat{\beta}_{X|C}$ to be positively biased; and vice versa for $R^2_\alpha < 0$. (See the Supplementary Materials for details of the alignment function.)

Program *causalsens* outputs a line plot and a table of $\hat{\beta}_{X|C,U(\phi)}$ and corresponding 95% CI for different values of R^2_α. Additionally, *causalsens* reports benchmarks for R^2_α based on the partial R^2 values for each covariate in C.

4.3 sensemakr

Program *sensemakr* uses formulae to estimate $\hat{\beta}_{X|C,U(\phi)}$ and its t-value for prespecified values of ϕ. Additionally, *sensemakr* reports summary measures, called "robustness values", which quantify the minimum amount of unmeasured confounding needed to change a study's conclusions, conditional on C[63].

The bias model of *sensemakr* expresses the absolute difference between the naive and bias-adjusted estimates, $\widehat{\Delta}_\phi = |\hat{\beta}_{X|C} - \hat{\beta}_{X|C,U(\phi)}|$, and the standard error of $\hat{\beta}_{X|C,U(\phi)}$ as functions of estimated quantities from the naive analysis and $\phi = (R^2_{X\sim U|C}, R^2_{Y\sim U|X,C})$. Bias parameter $R^2_{X\sim U|C}$ is the proportion of the variance of X, not explained by C, that is explained by U, and $R^2_{Y\sim U|X,C}$ is the proportion of the variance of Y, not explained by X and C, that is explained by U. Considering both directions of effect of U, $\hat{\beta}_{X|C,U(\phi)} = \hat{\beta}_{X|C} \pm \widehat{\Delta}_\phi$, and the corresponding t-value for a null hypothesis is $\frac{\hat{\beta}_{X|C,U(\phi)}}{se(\hat{\beta}_{X|C,U(\phi)})}$.

The robustness value for the point estimate (or t-value) represents the minimum value of $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$, when $R^2_{X\sim U|C} = R^2_{Y\sim U|X,C}$, such that $\hat{\beta}_{X|C,U(\phi)}$ (or its t-value) equals its prespecified tipping point value; for example, the null (or the 5% critical t-value). A robustness value close to 1 indicates that strong unmeasured confounding would be needed to change the study conclusions, whilst a value close to 0 indicates that very weak unmeasured confounding could change the conclusions.

Program *sensemakr* calculates upper bounds (called "benchmark bounds") for $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$ using C[63,70]. The benchmark bounds based on covariate C_j represent the maximum values for $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$ if U was k times $(k = 1, 2, 3, \ldots)$ as strong as C_j (in terms of strengths of relationships with X and Y)[63]. Additionally, *sensemakr* can calculate benchmark bounds based on a group of measured covariates.

Program *sensemakr* outputs robustness values for the point estimate and t-value, and contour plots of $\hat{\beta}_{X|C,U(\phi)}$ and corresponding t-value for prespecified values of ϕ, indicating the combinations of $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$ that correspond to a tipping point for the point estimate or t-value. Also, *sensemakr* outputs a table of benchmark bounds and values of $\hat{\beta}_{X|C,U(\phi)}$ and corresponding CI when $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$

equals these benchmark bounds. Note that, only the R package and Stata command can calculate benchmark bounds based on more than one measured covariate.

4.4 EValue

Program *EValue* reports a summary measure, called an E-value, which quantifies the minimum amount of unmeasured confounding needed to change a study's conclusions, conditional on the measured covariates[48]. The E-value is defined on the risk ratio scale and is a function of estimated quantities from the naive analysis and two bias parameters $\phi = (RR_{XU}, RR_{UY})$. For binary X and a single, binary U, RR_{XU} represents the risk ratio for the effect of X on U conditional on C and RR_{UY} represents the maximum risk ratio for the effect of U on Y after adjustment for C among the exposed and unexposed[78]. (See Ding and VanderWeele[78] for a definition of RR_{XU} and RR_{UY} when U denotes a single or multiple unmeasured confounders of type continuous, categorical or mixed.) For effect measures other than the risk ratio, the naive results are first converted to the risk ratio scale before calculating the E-value[48]. For example, for standardised mean difference, $\hat{\beta}^{std}_{X|C}$, and corresponding standard error, $SE_{\hat{\beta}^{std}_{X|C}}$, the approximate risk ratio for the point estimate is $\exp\{0.91 \times \hat{\beta}^{std}_{X|C}\}$ and the approximate risk ratio for a limit of the 95% CI is $\exp\{0.91 \times \hat{\beta}^{std}_{X|C} \pm 1.78 \times SE_{\hat{\beta}^{std}_{X|C}}\}$[48]. Note that, the E-value is interpreted on the risk ratio scale for all types of effect measures.

Here we describe the E-value when the tipping point of the point estimate is the null, although it can also be set to a non-null value (see Supplementary Materials of[48]). A separate E-value is calculated for the point estimate and CI limit closest to the null. The E-value for the point estimate (or CI limit) represents the minimum value of RR_{XU} and RR_{UY}, when $RR_{XU} = RR_{UY}$, such that $\hat{\beta}_{X|C,U(\phi)}$ is null or in the reverse direction to that of $\hat{\beta}_{X|C}$ (or the exposure effect is no longer statistically significant after adjustment for C and U). 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. When $\hat{\beta}_{X|C}$ is null (or its CI includes the null) the E-value for the point estimate (or CI limit) is 1, indicating that no unmeasured confounding is required to change the study conclusions. Importantly, the E-value is a measure of sensitivity to unmeasured confounding for a worst-case scenario (i.e., bias parameters RR_{XU} and RR_{UY} are set to values which maximize the bias due to unmeasured confounding)[79].

The R package *EValue*, Stata command *evalue*, and web tool *e-value calculator* can all be applied when the effect measure of interest is a risk ratio, risk difference, standardised mean difference, odds ratio or hazard ratio for a rare outcome (i.e., prevalence $< 15\%$), and odds ratio or hazard ratio for a common outcome (i.e., prevalence $\geq 15\%$). From here onward, we shall use *EValue* to represent all three implementations. Program *EValue* outputs E-values for the point estimate and CI limit, and a line plot of the values of RR_{UY} and RR_{XU} that correspond to prespecified tipping points for the point estimate and CI limit. Note that, the program does not supply benchmark values for RR_{UY} and RR_{XU}. For comparison purposes,

VanderWeele and Ding[48] suggest omitting each measured covariate in turn and recalculating the E-value.

4.5 konfound

Program *konfound* assesses sensitivity to a change in the statistical (in)significance status of $\hat{\beta}_{X|C}$[58]. This includes the scenario where U explains away 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). Program *konfound* refers to the first scenario as U "invalidating inference" and the second as U "sustaining inference". By default, the significance level is 5% and the null hypothesis is "no exposure effect", both of which can be changed by the analyst.

Program *konfound* reports two summary measures that quantify the minimum 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[80,81]. The formula for the percent bias is a function of estimated quantities from the naive analysis and the value of $\hat{\beta}_{X|C,U(\phi)}$ when its P-value is exactly $\kappa\%$ (for statistical significance defined at the $\kappa\%$ level). The impact threshold is also derived from estimated quantities of the naive analysis plus two bias parameters $\phi = (r_{X\sim U|C}, r_{Y\sim U|C})$: $r_{X\sim U|C}$ and $r_{Y\sim U|C}$ represent the partial correlation between U and X and between U and Y (conditional on C), respectively[82]. The impact threshold is the product $r_{X\sim U|C} \times r_{Y\sim U|C}$ when $r_{X\sim U|C}$ and $r_{Y\sim U|C}$ are equal and set to their minimum value such that statistical inference is invalidated or sustained. Note that, the percent bias measure is always positive but the impact threshold measure can be positive or negative depending on the direction of the correlation between U and X and Y. For both measures, larger absolute values indicate greater robustness to unmeasured confounding. Program *konfound* calculates the impact threshold and percent bias when Y is a continuous outcome and the naive analysis is a linear regression.

The software 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"; generated by the R package and online tool). Only the Stata command provides benchmark values for $r_{X\sim U|C}$ and $r_{Y\sim U|C}$, which are the partial correlation of each measured covariate C_j with X and with Y, respectively, given the remaining measured covariates.

5 Illustrative example

We applied the five QBA methods of Section 4 to data from the BCG and NHANES studies. For 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. In the BCG example, U was a single confounder and

adjustment for U did not change the study conclusions. See the Supplementary Materials for the NHANES example where U represents multiple confounders.

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 C, we calculated benchmark E-values and, for the other four programs, we calculated benchmark values for ϕ and the bias-adjusted results when ϕ was set to (multiples of) the benchmark values corresponding to the "strongest measured covariate" (i.e., the covariate that had the strongest associations with X and Y).

As this is an illustrative example of applying a 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.

5.1 Description of the BCG Study

The BCG study is a follow-up of a dietary intervention randomized controlled trial of pregnant women and their offspring[27,28]. 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 ≥ 17.44 kg/m^2 [83]). Measured covariates C were participant's sex, gestational age, birth weight, and parents' height and weight measurements. The strongest measured covariate was maternal weight. The unmeasured confounder U was a measure of childhood socioeconomic position (SEP) (paternal occupational social class based on the UK registrar general classification[84]) with $U = 1$ for professional or managerial occupations, and $U = 0$ otherwise. Based on the 542 participants with complete data on all variables, $\hat{\beta}_{X|C}$ was 2.21 kg/m^2 (95% CI 1.30, 3.11 kg/m^2) and the fully adjusted estimate (i.e., adjusted for C and U) was 2.19 kg/m^2 (95% CI $1.29, 3.09$ kg/m^2). Also, the coefficient of U from the linear regression $Y|X, C, U$ was -0.66 kg/m^2 (95% CI $-1.57, 0.25$ kg/m^2) and the coefficient of U from the logistic regression $X|C, U$ was -0.23 (95% CI $-0.85, 0.35$). Statistical significance was defined at the 5% level.

Note that, on a computer with 2.7 Ghz the run-time of *treatSens* (with the default setting of single-threading[46]) was 10 minutes while the other programs generated their results instantaneously. We begin with a description of the outputted results and then compare the results across the five programs. In the Supplementary Materials we report on a small survey we conducted to obtain feedback on how the five QBA programs compare with respect to ease/difficultly of interpreting their QBA results.

5.2 Results

treatSens

Program *treatSens* outputs a contour plot (Figure 1(a)) where each contour represents the different combinations of $\phi = (\zeta^Y, \zeta^Z)$ that result in the same bias-adjusted estimate, $\hat{\beta}_{X|C,U(\phi)}$. For example, $\hat{\beta}_{X|C,U(\phi)} = 0.40$ standard deviations of BMI (SD-BMI; or equivalently $\hat{\beta}_{X|C,U(\phi)} = 1.81$ kg/m^2) when $\zeta^Y = 0.23$ and $\zeta^Z = 1.00$, and when $\zeta^Y = 1.00$ and $\zeta^Z = 0.24$. (Note that, *treatSens* standardises all continuous variables.) The black horizontal contour at $\zeta^Y = 0$ denotes the naive estimate of 0.49 SD-BMI (i.e., $\hat{\beta}_{X|C} = 2.21$ kg/m^2), 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: pluses represent covariates positively associated with adult BMI, and the inverted triangles represent covariates 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).

causalsens

Program *causalsens* outputs a line plot (Figure 1(b)) where the black line represents the bias-adjusted exposure estimates, the grey shaded area represents the corresponding 95% CIs, and the crosses denote the benchmark values for $\phi = (R^2_\alpha)$ with each benchmark appearing twice to allow for both directions of effect. Values of $R^2_\alpha > 0$ implies that individuals in the unexposed group tended to be healthier (i.e., lower adult BMI) than those in the exposed group even if everyone was of normal weight (or overweight) at age 5; and the converse for $R^2_\alpha < 0$.

sensemakr

Program *sensemakr* outputs four contour plots: Figures 1(c) and (d) show the contour plots for the exposure effect estimate and its t-value, respectively, generated under the assumption that accounting for U moves the exposure effect estimate closer to the null, and Supplementary Figures S2(a) and (b) show the same contour plots generated under the converse assumption. The contours have a similar interpretation as discussed for *treatSens*. For example, the red contour represents different combinations of $\phi = (R^2_{X\sim U|C}, R^2_{Y\sim U|X,C})$ that result in a null exposure effect (Figure 1(c)) and the critical t-value corresponding to 5% statistical significance (Figure 1(d)). The black triangle denotes the naive estimate, $\hat{\beta}_{X|C}$, and the red diamonds denote once, twice and thrice the benchmark bounds based on the strongest measured covariate.

The robustness values for $\hat{\beta}_{X|C}$ and its t-value were 18.76% and 11.56%, respectively. Thus, U would need to explain at least 18.76% (or 11.56%) of the residual variance of both childhood overweight and adult BMI for the exposure effect adjusted for C and U to be null (or statistically insignificant).

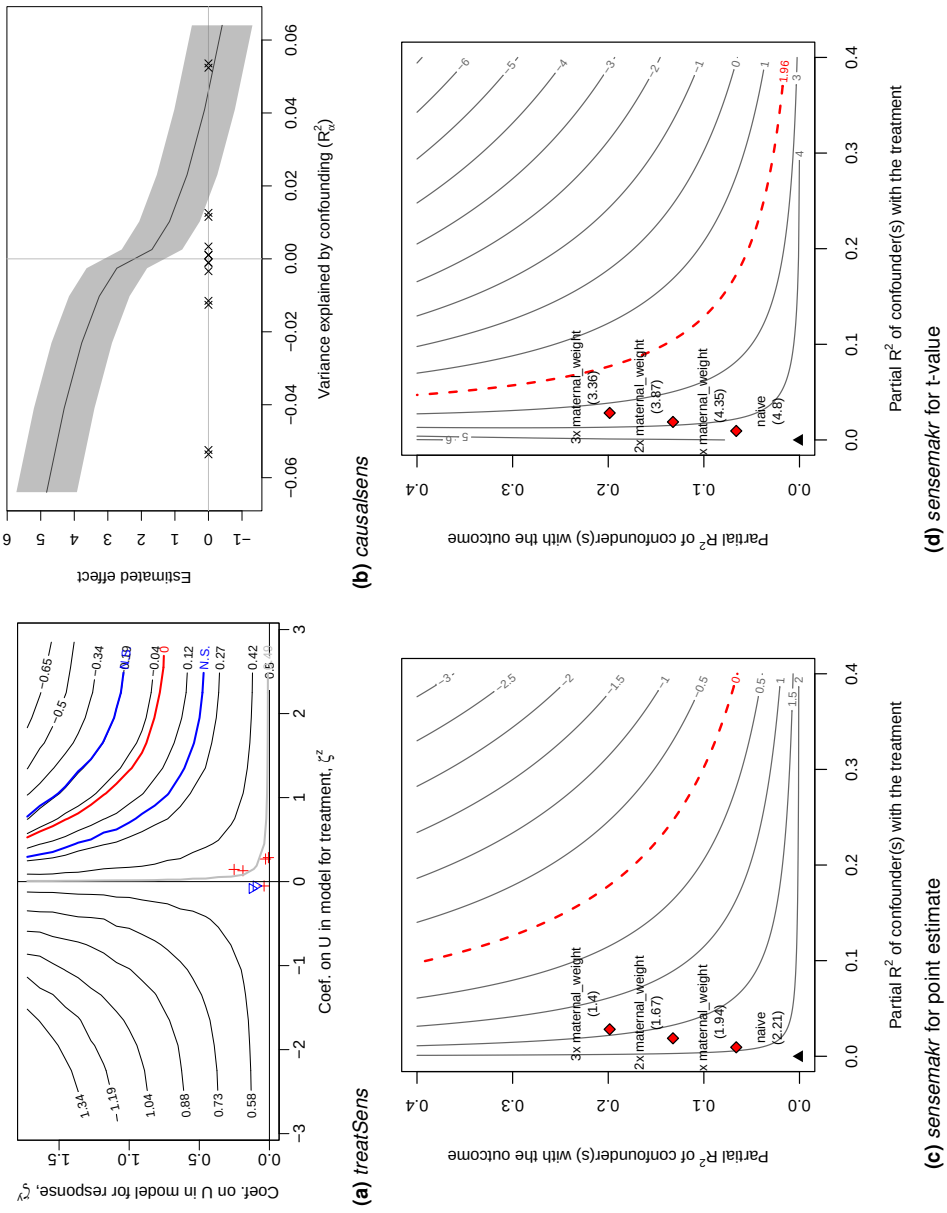
EValue

The E-values for $\hat{\beta}_{X|C}$ and its lower CI limit were 2.50 and 1.93, respectively. Thus, if the associations between U and adult BMI and childhood overweight were at least 2.50

(or 1.93), on the risk ratio scale, then the exposure effect adjusted for C and U may be null or in the reverse direction (or strictly positive but statistically insignificant). Supplementary Figure S3 shows the combinations of $\phi = (RR_{UY}, RR_{XU})$ that correspond to a null bias-adjusted estimate (red contour) and a strictly positive but statistically insignificant bias-adjusted estimate (black contour).

konfound

The percent bias was 59.11%, depicted in the bar-graph shown in Supplementary Figure S4, and the impact threshold was 0.13 with bias parameters $r_{X\sim U|C} = r_{Y\sim U|C} = \sqrt{0.13}$, depicted in the causal diagram shown in Supplementary Figure S5. Therefore, in order for the exposure effect to be statistically insignificant after adjustment for C and U then (1) U would need to account for at least 59.11% of $\hat{\beta}_{X|C}$ (i.e., $\hat{\beta}_{X|C,U(\phi)} \leq 0.90$ kg/m^2), and (2) the partial correlations of U with adult BMI and child overweight must both exceed 0.36.



(a) *treatSens* **(b)** *causalsens* **(c)** *sensemakr* for point estimate **(d)** *sensemakr* for t-value

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).

Comparison of the results

Table 2 summarises the bias-adjusted results of each program in scenarios where the associations between U and adult BMI and child overweight were half, once and twice as strong as the corresponding associations with the strongest measured covariate (i.e., ϕ set to $0.5, 1$ and $2 \times$ benchmark values for maternal weight).

Table 2. Bias-adjusted estimate $\hat{\beta}_{X|C,U(\phi)}$ and corresponding 95% confidence interval when bias parameter ϕ is set to specified multiples of benchmark values based on strongest measured confounder maternal weight (MW)

	$\hat{\beta}_{X\|C,U(\phi)}$ [95% confidence interval] in kg/m^2			
If ϕ set to	*treatSens*	*causalsens*	*sensemakr*	*konfound*
Bias towards the null				
$0.5\times$ benchmark values of MW	2.19 [1.29, 3.09]	0.52 [−0.39, 1.42]	2.08 [1.18, 2.97]	[excludes 0]
$1\times$ benchmark values of MW	2.14 [1.24, 3.04]	−0.19 [−1.09, 0.72]	1.94 [1.06, 2.82]	[excludes 0]
$2\times$ benchmark values of MW	1.91 [1.04, 2.78]	−1.18 [−2.08, −0.28]	1.67 [0.82, 252]	[excludes 0]
Bias away from the null				
$0.5\times$ benchmark values of MW	2.24 [1.34, 3.14]	3.90 [3.00, 4.80]	2.34 [1.45, 3.23]	[excludes 0]
$1\times$ benchmark values of MW	2.31 [1.42, 3.21]	4.60 [3.70, 5.51]	2.47 [1.60, 3.35]	[excludes 0]
$2\times$ benchmark values of MW	2.54 [1.67, 3.41]	5.59 [4.69, 6.50]	2.74 [1.89, 3.59]	[excludes 0]

Considering unmeasured confounding towards or away from the null, if U was comparable to the strongest measured covariate (with respect to its associations with adult BMI and child overweight) then *treatSens* and *sensemakr* report that adjusting for C and U would give similar results to those of the naive analysis and *konfound* indicates the exposure effect would remain statistically significant. Also, *sensemakr*'s robustness values were substantially higher than the benchmark bounds for $R^2_{X\sim U|C}$ and $R^2_{Y\sim U|X,C}$ even when these benchmarks were based on all of C (Supplementary Table S1). Similarly, the benchmark E-values when omitting the strongest measured covariate and U were comparable to the E-values when omitting U only (Supplementary Table S2), indicating that the exposure effect adjusted for C and U would remain above the null and statistically significant. Furthermore, *treatSens*, *sensemakr*, and *konfound* indicate that U would need to be more than double the strength of the strongest measured covariate in order to change the study conclusions (i.e., a null or doubling of the exposure effect, or a statistically insignificant effect). Conversely, *causalsens* suggests adjusting for U comparable to the strongest measured covariate could result in an exposure effect close to the null or more than double the naive estimate.

Provided the naive analysis included all of the important confounders then it seems unlikely that the confounding effect of U, childhood SEP, could be more than twice as strong as the strongest measured covariate, especially given that childhood SEP would likely be correlated with at least some of the measured covariates. Therefore, under these assumptions, *treatSens*, *sensemakr*, *konfound*, and *EValue* indicates robustness of the BCG study conclusions to unmeasured confounding by childhood SEP which was inline with the fully adjusted results. In contrast, *causalsens* suggested study conclusions could differ if we were able to adjust for childhood SEP.

6 Discussion

We have conducted an up-to-date review of software implementations of QBA to unmeasured confounding, and a detailed illustration of the latest software applicable for a linear regression analysis of an unmatched study. All programs implement a deterministic QBA, and most are available in the free software environment R. The majority were developed in the latter half of the past decade and include programs available when the naive analysis is a mediation analysis, meta-analysis and a survival analysis. Many programs include features such as benchmarking and graphical displays of the QBA results to aid interpretation. Our comparative example illustrated that even QBA software applicable to the same naive analysis can implement distinct QBA methods. All programs were straightforward to implement and instantly generated the results except for *treatSens* which took about 10 minutes to run when applied to a moderately-sized dataset. All programs provided information about the amount of unmeasured confounding at the tipping points; however, *treatSens*, *sensemakr* and *causalsens* also provided information on the bias-adjusted results for any specified level of unmeasured confounding with minimal extra burden to the analyst.

Out of the five programs we compared *sensemakr* performs the most detailed QBA. It generates bias-adjusted results for prespecified levels of unmeasured confounding (similarly to *treatSens* and *causalsens*), reports a summary measure at prespecified

tipping points (similarly to *EValue* and *konfound*) and conducts a QBA in a worse-case scenario of unmeasured confounding (similarly to *EValue*). However, in our small panel study, three out of seven participants reported difficulties interpreting the output of *sensemakr*. Program *EValue* implements a flexible QBA which can be applied to a wide range of effect measures and makes minimal assumptions about the unmeasured confounding (e.g., allows U to be a modifier of the $X-Y$ relationship). However, the downside of this flexibility is that the analyst may be unaware of the additional assumptions required when converting their effect measure to the risk ratio scale and it can be challenging to establish plausible values for its bias parameters (either from external data or from benchmarking). Also, a notable limitation of programs *EValue* and *konfound* is that they are restricted to establishing robustness to unmeasured confounding (i.e., cannot provide results adjusted for likely levels of unmeasured confounding) and *konfound* only considers sensitivity to changes in statistical significance. The upside of the programs' simplicity is that they require only summary data and so can be easily applied to multiple published studies, with the *EValue* extended to random-effects meta analyses[51]. Three strengths of *treatSens* over the other programs are: (1) its imputation-style QBA method will be familiar to many analysts, (2) its bias parameters (i.e., regression coefficients) are more likely to be reported by published studies than the bias parameters of the other programs (e.g., partial R^2 values), and (3) *treatSens* can also be applied when the analysis of interest is a non-parametric model (Bayesian additive regression tree). A potential weakness of *treatSens* is that it simulates U from a limited choice of joint distributions.

A limitation of our review is that we focused on software described in the published literature. We recognise that additional software programs are available such as other implementations of QBA methods discussed in this review (e.g., another implementation of the E-value[85]) and programs of other QBA methods (e.g., TippingSens[86]). Our illustrative example compared software programs applicable when the analysis of interest is a linear regression since previous comparisons of QBA methods have primarily focused on analyses of binary outcomes[10,19–26]. Of the software we compared, programs *konfound* and *EValue* can be applied to a binary outcome, with *EValue* also applicable when the exposure effect is a hazard ratio. Future work could compare QBA methodology for analyses of other types of outcomes such as survival and categorical outcomes.

Several programs in our review provided benchmark values to aid interpretation of the QBA results. Note that, *sensemakr* can provide benchmark bounds for its bias parameters based on a group of measured covariates which provides a useful aid when considering multiple unmeasured confounders. Interestingly, participants of our small panel study reported difficulties interpreting the E-value in the absence of any benchmarks. One noted issue with benchmarking is that the benchmarks tend to be based on the naive models, $Y|X,C$ and $X|C$, and do not adjust for the omission of U[30,63]. See Cinelli and Hazelett for a discussion on why ignoring U can affect the benchmark estimates even when U is assumed to be independent of C[63]. Examples of QBAs using benchmarking that accounts for the omission of U include *sensemakr*,[30], and[87].

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[63]. Alternatively, a QBA method may leave the functional form of U unspecified and instead define its bias parameters as upper bounds (such as the *EValue* 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[78]). 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[88], 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. We also thank researchers at the Leiden University Medical Center for participating in our panel study.

Funding

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 and Robins J. *Causal inference: What if*. 1 ed. Boca Raton: Chapman & Hill/CRC, 2020.
2. Arah OA. Bias analysis for uncontrolled confounding in the health sciences. *Annu Rev Public Health* 2017; 38: 23–38.
3. Fewell Z, Davey Smith G and Sterne JAC. The impact of residual and unmeasured confounding in epidemiological studies: a simulation study. *Am J Epidemiol* 2007; 166(6): 646–655.
4. Groenwold RH, Sterne JA, Lawlor DA et al. Sensitivity analysis for the effects of multiple unmeasured confounders. *Ann Epidemiol* 2016; 26(9): 605–611.
5. Uddin MJ, Groenwold RH, Ali MS et al. Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *Int J Clin Pharm* 2016; 38(3): 714–723.
6. Pouwels KB, Widyakusuma NN, Groenwold RH et al. Quality of reporting of confounding remained suboptimal after the strobe guideline. *J Clin Epidemiol* 2016; 69: 217–224.
7. Lash TL, Fox MP and Fink AK. *Applying quantitative bias analysis to epidemiologic data*. 1 ed. New York: Springer, 2009.
8. Lash TL, Fox MP, MacLehose RF et al. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014; 43(6): 1969–1985.
9. Hunnicutt JN, Ulbricht CM, Chrysanthopoulou SA et al. Probabilistic bias analysis in pharmacoepidemiology and comparative effectiveness research: a systematic review. *Pharmacoepidemiol Drug Saf* 2016; 25(12): 1343–1353.
10. Liu W, Kuramoto SJ and Stuart EA. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prev Sci* 2013; 14(6): 570–580.
11. Peel MJ. Addressing unobserved endogeneity bias in accounting studies: control and sensitivity methods by variable type. *Account Bus Res* 2014; 44(5): 545–571.
12. Streeter AJ, Lin NX, Crathorne L et al. Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *J Clin Epidemiol* 2017; 87: 23–34.
13. Budziak J and Lempert D. Assessing threats to inference with simultaneous sensitivity analysis: the case of us supreme court oral arguments. *Political Sci Res Methods* 2018; 6(1): 33–56.
14. Zhang X, Faries DE, Li H et al. Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiol Drug Saf* 2018; 27(4): 373–382.
15. Zhao Q, Small DS and Bhattacharya BB. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *J R Stat Soc Series B Stat Methodol* 2019; 81(4): 735–761.

16. Barberio J, Ahern TP, MacLehose RF et al. Assessing techniques for quantifying the impact of bias due to an unmeasured confounder: an applied example. *Clin Epidemiol* 2021; 13: 627–635.
17. Qin X and Yang F. Simulation-based sensitivity analysis for causal mediation studies. *Psychol Methods* ; Epub ahead of print 16 December 2021. DOI: 10.1037/met0000340.
18. Rosenbaum PR. *Observational Studies*. 2 ed. New York: Springer, 2002.
19. Arah OA, Chiba Y and Greenland S. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. *Ann Epidemiol* 2008; 18: 637–646.
20. Groenwold RH, Nelson DB, Nichol KL et al. Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. *Int J Epidemiol* 2010; 39(1): 107–117.
21. MacLehose RF, Ahern TP, Lash TL et al. The importance of making assumptions in bias analysis. *Epidemiol* 2021; 32(5): 617.
22. McCandless LC and Gustafson P. A comparison of Bayesian and Monte Carlo sensitivity analysis for unmeasured confounding. *Stat Med* 2017; 36(18): 2887–2901.
23. Mittinty MN. Estimation bias due to unmeasured confounding in oral health epidemiology. *Community Dent Health* 2020; 37: 1–6.
24. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf* 2006; 15(5): 291–303.
25. Steenland K and Greenland S. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *Am J Epidemiol* 2004; 160(4): 384–392.
26. Thommes EW, Mahmud SM, Young-Xu Y et al. Assessing the prior event rate ratio method via probabilistic bias analysis on a Bayesian network. *Stat Med* 2020; 39(5): 639–659.
27. Elwood PC, Haley T, Hughes S et al. Child growth (0-5 years), and the effect of entitlement to a milk supplement. *Arch Disin Child* 1981; 56(11): 831–835.
28. McCarthy A, Hughes R, Tilling K et al. Birth weight; postnatal, infant, and childhood growth; and obesity in young adulthood: evidence from the Barry Caerphilly Growth study. *Am J Clin Nutr* 2007; 86(4): 907–913.
29. Centers for Disease Control and Prevention/National Center for Health Statistics. National Health and Nutrition Examination Survey data. `https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2015` (2016, accessed 17 October 2022).
30. Zhang B and 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. *J R Stat Soc C: Appl Stat* 2020; 69(5): 1285–1305.
31. 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/`.
32. Rosenbaum PR. Hodges-Lehmann point estimates of treatment effect in observational studies. *J Am Stat Assoc* 1993; 88(424): 1250–1253.
33. Harada M. ISA: Stata module to perform Imbens'(2003) sensitivity analysis. `https://econpapers.repec.org/software/bocbocode/s457336.htm`, (2012, accessed on 17 October 2022).
34. Imbens GW. Sensitivity to exogeneity assumptions in program evaluation. *Am Econ Rev* 2003; 93(2): 126–132.

35. Harada M. GSA: Stata module to perform generalized sensitivity analysis. https://econpapers.repec.org/software/bocbocode/s457497.htm, (2012, accessed on 17 October 2022).
36. Small DS, Cheng J, Halloran ME et al. Case definition and design sensitivity. *J Am Stat Assoc* 2013; 108(504): 1457–1468.
37. Small D. Sensitivitycasecontrol: Sensitivity analysis for case-control studies. https://cran.r-project.org/web/packages/SensitivityCaseControl/SensitivityCaseControl. (2015, accessed on 17 October 2022).
38. Blackwell M. A selection bias approach to sensitivity analysis for causal effects. *Polit Anal* 2014; 22(2): 169–182.
39. Blackwell M. causalsens: Selection bias approach to sensitivity analysis for causal effects. https://cran.r-project.org/web/packages/causalsens/causalsens.pdf, (2018, accessed on 17 October 2022).
40. Subramanian HC and Overby E. mbsens: module to compute sensitivity metric for matched sample using mcnemar's test. https://econpapers.repec.org/software/bocbocode/s457867.htm, (2014, accessed on 17 October 2022).
41. Rosenbaum PR. Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies. *Biometrics* 2009; 63: 456–464.
42. Rosenbaum PR. Two r packages for sensitivity analysis in observational studies. *Observational Studies* 2015; 1(2): 1–17.
43. Rosenbaum PR. sensitivitymw: Sensitivity analysis using weighted M-statistics. https://CRAN.R-project.org/package=sensitivitymw, (2015, accessed on 17 October 2022).
44. Carnegie NB, Harada M and Hill JL. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *J Res Edu Eff* 2016; 9(3): 395–420.
45. Dorie V, Harada M, Carnegie NB et al. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Stat Med* 2016; 35(20): 3453–3470.
46. Carnegie NB, Harada M, Dorie V et al. treatsens: Sensitivity analysis for causal inference. https://mran.microsoft.com/snapshot/2018-03-11/web/packages/treatSens/treatSens.pdf, (2018, accessed on 17 October 2022).
47. Rosenbaum PR. sensitivitymv: Sensitivity analysis in observational studies. https://CRAN.R-project.org/web/packages/sensitivitymv/sensitivitymv.pdf, (2018, accessed on 17 October 2022).
48. VanderWeele TJ and Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med* 2017; 167(4): 268–274.
49. Mathur MB, Ding P, Riddell CA et al. Website and R package for computing E-values. *Epidemiol* 2018; 29(5): e45.
50. Linden A, Mathur MB and VanderWeele TJ. Conducting sensitivity analysis for unmeasured confounding in observational studies using E-values: The evalue package. *SJ* 2020; 20(1): 162–175.
51. Mathur MB and VanderWeele TJ. Sensitivity analysis for unmeasured confounding in meta-analyses. *J Am Stat Assoc* 2020; 115(529): 163–172.
52. Hong G, Qin X and Yang F. Weighting-based sensitivity analysis in causal mediation studies. *J Educ Behav Stat* 2018; 43(1): 32–56.

53. Qin X, Hong G and Yang F. rmpw: Causal mediation analysis using weighting approach. https://cran.r-project.org/web/packages/rmpw/rmpw.pdf, (2018, accessed on 17 October 2022).
54. Rosenbaum PR. sensitivityfull: Sensitivity analysis for full matching in observational studies. https://CRAN.R-project.org/web/packages/sensitivityfull/sensitivityfull.pdf, (2017, accessed on 17 October 2022).
55. Aikens RC, Greaves D and Baiocchi M. A pilot design for observational studies: Using abundant data thoughtfully. *Stat Med* 2020; 39: 4829–4840.
56. Lee K, Small DS and Rosenbaum PR. A powerful approach to the study of moderate effect modification in observational studies. *Biometrics* 2018; 74(4): 1161–1170.
57. Lutz SM, Thwing A, Schmiege S et al. Examining the role of unmeasured confounding in mediation analysis with genetic and genomic applications. *BMC Bioinform* 2017; 18(1): 1–6.
58. Xu R, Frank KA, Maroulis SJ et al. konfound: Command to quantify robustness of causal inferences. *SJ* 2019; 19(3): 523–550.
59. Rosenberg JM, Xu R and Frank KA. KonFound-It!: Quantify the robustness of causal inferences. https://CRAN.R-project.org/web/packages/konfound/konfound.pdf, (2021, accessed on 17 October 2022).
60. Lindmark A, de Luna X and Eriksson M. Sensitivity analysis for unobserved confounding of direct and indirect effects using uncertainty intervals. *Stat Med* 2018; 37(10): 1744–1762.
61. Lindmark A. sensmediation: Parametric estimation and sensitivity analysis of direct and indirect effects. https://cran.r-project.org/web/packages/sensmediation/sensmediation.pdf, (2019, accessed on 17 October 2022).
62. Zhang B. sensitivitycalibration: A calibrated sensitivity analysis for matched observational studies. https://CRAN.R-project.org/web/packages/sensitivityCalibration/sensitivityCalibra (2018, accessed on 17 October 2022).
63. Cinelli C and Hazlett C. Making sense of sensitivity: Extending omitted variable bias. *J R Stat Soc Ser B Methodol* 2020; 82(1): 39–67.
64. Cinelli C, Ferwerda J and Hazlett C. sensemakr: Sensitivity analysis tools for regression models. https://www.researchgate.net/publication/340965014_sensemakr_Sensitivity_Analysis_ (2020, accessed on 17 October 2022).
65. Genbäck M and de Luna X. Causal inference accounting for unobserved confounding after outcome regression and doubly robust estimation. *Biometrics* 2019; 75(2): 506–515.
66. Qin X and Yang F. mediationsens: Simulation-based sensitivity analysis for causal mediation. https://cran.r-project.org/web/packages/mediationsens/mediationsens.pdf, (2020, accessed on 17 October 2022).
67. Huang R, Xu R and Dulai PS. Sensitivity analysis of treatment effect to unmeasured confounding in observational studies with survival and competing risks outcomes. *Stat Med* 2020; 39(24): 3397–3411.
68. Huang R. survsens: Sensitivity analysis with time-to-event outcomes. https://CRAN.R-project.org/web/packages/survSens/survSens.pdf, (2020, accessed on 17 October 2022).
69. Liu X and Wang L. The impact of measurement error and omitting confounders on statistical inference of mediation effects and tools for sensitivity analysis. *Psychol Methods* 2021; 26(3): 327–342.

70. Cinelli C, Kumor D, Chen B et al. Sensitivity analysis of linear structural causal models. In *International Conference on Machine Learning*. California: PMLR, pp. 1252–1261.
71. Cinelli C, Ferwerda J, Hazlett C et al. sensemakr: Sensitivity analysis tools for regression models. `https://cran.r-project.org/web/packages/sensemakr/sensemakr.pdf`, (2021, accessed on 17 October 2022).
72. Mathur MB, Smith LH, Ding P et al. Evalue: Sensitivity analysis for unmeasured confounding and other biases in observational studies and meta-analyses. `https://cran.r-project.org/web/packages/EValue/EValue.pdf`, (2021, accessed on 17 October 2022).
73. Rosenberg JM, Xu R, Lin Q et al. KonFound-It!: Quantify the robustness of causal inferences. `http://konfound-it.com`, (2022, accessed on 17 October 2022).
74. Rubin D. Multiple imputation after 18+ years. *J Am Stat Assoc* 1996; 91(434): 473–489.
75. Rubin DB. Causal inference using potential outcomes. *J Am Stat Assoc* 2005; 100(469): 322–311.
76. Robins JM. Association, causation and marginal structural models. *Synthese* 1999; 121: 151–179.
77. Robins JM. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models, section 6–11. In Halloran M and Berry D (eds.) *Statistical Models in Epidemiology: The Environment and Clinical Trials*. Springer–Verlag: New York, 1999.
78. Ding P and VanderWeele TJ. Sensitivity analysis without assumptions. *Epidemiol* 2016; 27(3): 368.
79. VanderWeele TJ. Are greenland, ioannidis and poole opposed to the cornfield conditions? a defence of the e-value. *Int J Epidemiol* 2022; 51(2): 364–371.
80. Frank K and Min KS. Indices of robustness for sample representation. *Sociol Methodol* 2007; 37(1): 349–392.
81. Frank KA, Maroulis SJ, Duong MQ et al. What would it take to change an inference? using Rubin's causal model to interpret the robustness of causal inferences. *Educ Eval Policy Anal* 2013; 35(4): 437–460.
82. Frank KA. Impact of a confounding variable on a regression coefficient. *Sociol Methods Res* 2000; 29(2): 147–194.
83. Cole TJ, Bellizzi MC, Flegal KM et al. Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ* 2000; 320(7244): 1240.
84. Pevalin D and 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. *Soc Contemp* 2002; (1): 75–106.
85. Haine D. Compute e-value to assess bias due to unmeasured confounder. `https://dhaine.github.io/episensr/reference/confounders.evalue.html#references`, (2018, accessed on 17 October 2022).
86. Haensch AC, Drechsler J and Bernhard S. Tippingsens: An r shiny application to facilitate sensitivity analysis for causal inference under confounding. `https://www.econstor.eu/bitstream/10419/234287/1/dp2029.pdf`, (2018, accessed on 17 October 2022).
87. Hsu JY and Small DS. Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics* 2013; 69(4): 803–811.

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