Community Dental Health (2020) **37,** 84–89 © BASCD 2020
doi:10.1922/CDH_SpecialIssueMittinty06

Estimating Bias Due to Unmeasured Confounding in Oral Health Epidemiology

[Special issue of Community Dental Health, to be disseminated at 'Random and Systematic Bias in Population Oral Health Research' IADR symposium, March 2020, Washington DC.]

Murthy N Mittinty[1,2]

[1]*School of Public Health, University of Adelaide, 57 North Terrace, AHMS building Level 9, Australia, 5000.*
[2]*Robinson Research Institute, University of Adelaide, North Adelaide, Australia 5000.*

Abstract: Confounding can make an association seem bigger when the true effect is smaller or vice-versa and it can also make it appear negative when it may actually be positive. In short, both the direction and the magnitude of an association are dependent on confounding. Therefore, understanding and adjusting for confounding in epidemiological research is central to addressing whether an observed association is causal or not. Moreover, unmeasured confounding in observational studies can give rise to biased estimates. Several techniques have been developed to account for bias and conducting sensitivity analysis. Using an hypothetical example this paper illustrates application of simple methods for conducting sensitivity analysis for unmeasured confounder(s).

Keywords: *Unmeasured confounding, bias, sensitivity, oral-health.*

Introduction

Consider scenario 1, comparing the incidence rate of tooth loss in children within a community in the year before and after the introduction of water supply fluoridation. The difference measured across different time points is *not* an *effect measure* but a *measure of association*, even though there may be a complete overlap of the children in both periods. So what exactly is the distinction between measure of effect and measure of association?

A *measure of effect* compares what would happen to one population under two possible but distinct conditions, where only one can occur (e.g., having cancer or not having cancer; we cannot observe both within the same person at the same time). Therefore the *measure of effect* is a theoretical concept insofar as it is logically impossible to observe the population under both the conditions at the same time; hence logically making it impossible to measure the effect directly. Conversely, a *measure of association* compares what happens in two distinct populations, although these populations may correspond to the same population in different time periods.

Given the observable nature of association measures, it is inviting to swap them for effect measures. It is even more natural to give causal explanations for observed associations in terms of obvious differences between populations being compared. However, this can be misleading. Let's look at Scenario 1 again and analyse in detail how a *measure of association* translates into a *measure of effect*. The desired effect to be measured is of fluoridation on tooth loss. To measure this effect, we must contrast the actual rate under fluoridation with the rate that would have occurred in the same period had fluoridation not been introduced. However, we cannot observe the latter because fluoridation was introduced, so the "*non fluoridation*" rate is the counterfactual. Suppose we exchange the rate in the time period before fluoridation, and also suppose that the rate before fluoridation equals the post fluoridation counterfactual rate; then the measure of association equals our desired measure of effect, and the before and after difference is un-confounded (i.e., confounding is not present in this difference). In other words, the exchangeability assumption ensures that the counterfactual risk (marginal risk estimate) under exposure to fluoridation ("yes", "no") equals the observed risk (conditional risk estimate) among those who received fluoridation. It is in this situation that the causal risk equals the associational risk. However, if the two differences are not equal, then the measure of association is not equal to the measure of effect for which it is substituted. In this circumstance, the measure of association is confounded (Greenland, 1996). It should be noted that this definition of confounding is not only applicable to differences but also to rate ratios.

Thus confounders can be defined as variables (e.g., exposures/treatments) that explain or produce part of the difference between the measure of association and the measure of effect that would be obtained with a counterfactual ideal (Hernan and Robins, 2020). Most epidemiological (including dental) research is concerned with adjusting or removing confounding. This is because confounding can both change the direction of the effect and also make a null effect causal or preventive (i.e., no causal relation between exposure and outcome) (Lash

Correspondence to: Murthy N Mittinty, Robinson Research Institute, University of Adelaide, North Adelaide, Australia 5000.
Email: murthy.mittinty@adelaide.edu.au

and Fink, 2003; Lash *et al*., 2011). For this reason, an analysis of causal effect of an exposure on the occurrence of an outcome (e.g., disease) must account for confounding of the crude association. When data on confounders is not available, then the investigator cannot control for the effect of confounding on study results through stratification or regression (Greenland. 1996). Then one can ask questions like "What impact might the uncontrolled confounding have had on the results, both regarding the direction of the uncontrolled confounding and its expected magnitude?" In both cases, bias analysis can be a useful tool to explore the impact of the unknown/unmeasured confounder. The aim of this paper is to illustrate how to use the tools and conduct bias analysis related to unmeasured confounders in oral health.

History and motivation for bias analysis due to unmeasured confounders

About half a century ago Cornfield (1959) and Bross (1966) proposed guidelines for determining whether an unmeasured binary covariate having specified properties could explain all of the apparent effect of treatment. That is, whether the treatment effect, after adjusting for unmeasured confounding, could be zero. Schlesselman (1978) developed methods to assess the effects of unmeasured confounding. Rosenbaum and Rubin (1983) presented methods for assessing the bias due to unobserved binary covariates on an outcome. Despite these advances, none of these methods is widely taught in undergraduate or graduate courses in biostatistics and epidemiology. Even research published in peer-reviwed journals does not acknowledge/describe/discuss unmeasured confounding, which perpetuates notion that there is seldom measurement error or unmeasured confounding in research. This complete absence of bias analysis may be due to lack of available software.

Observational studies are being used in etiological, prediction (e.g., identification of high-risk groups), prognostic and diagnostic research. Both clinical and public health researchers rely heavily on observational data. With the advent of large observational data from registries, electronic health record sample size is no longer an issue, thus reducing random error. However, the role of systematic errors/bias remains important in observational studies (Sterne *et al*., 2016). Unmeasured confounding is one such source of systematic error, hence assessing bias due to unmeasured confounding should be part of routine data analysis.

Notation:

Let's suppose that we have a data set, D, that comprises exposure, X, outcome, Y, and some measured confounders, C and unmeasured confounders U. That is $D = (C,X,Y,U)$. Usually U is not part of the observed D as they are unmeasured. Furthermore, let's suppose that the parameter of interest is either the risk difference (RD) or risk ratio or relative risk (RR). Additionally, supposing that we have a good understanding on how the data has been generated, we can illustrate it using a directed acyclic graph (DAG), mapping the relation between confounders, exposures and outcomes, as shown in Figure 1.

In Figure 1, circles denoting the nodes and arrows are known as the edges. You will note the arrows are pointing in one direction. C is a parent of both X and Y, Y is created by both X and C, and X is created by C. The confounder C can be a vector (e.g., age, gender…).

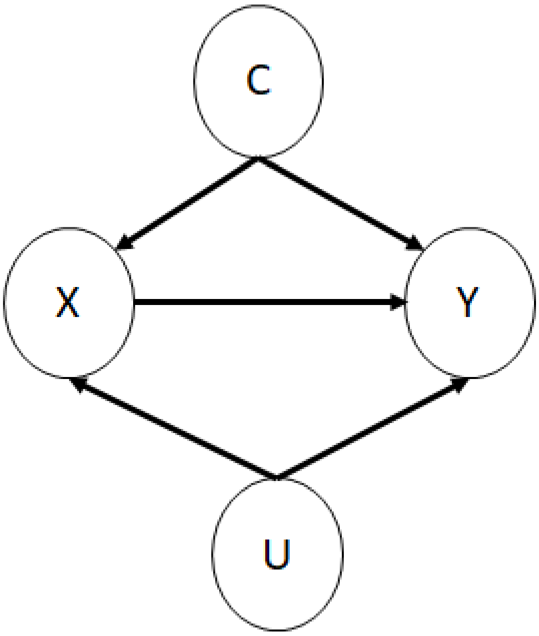

Figure 1: Directed acyclic graph mapping relations between confounders, exposures and outcomes.

Reasons for conducting bias analysis for unmeasured confounders

There could be several reasons, including:

1. Understanding whether an unmeasured confounder could explain the obtained results, i.e., to understand if the required values of bias parameters are plausible characteristics of an unmeasured confounder.
2. Understanding if a combination of bias parameters exists that would reverse the direction of an observed true effect, i.e., to check if specific combination of bias parameters would make a causal association appear as a preventive association or vice versa.

If the investigator's motivation is similar to (1), then the methods described in this paper are appropriate. However, if the interest is (2), then multidimensional techniques beyond the scope of this paper are required (Lash *et al.,* 2011).

Bias parameters

In order to conduct bias analysis for unmeasured confounders, the investigator must have knowledge of two parameters:

1. The association between the confounder and the outcome among those who were not exposed.
2. The prevalence of the unmeasured confounder in the source population.

Implementation of simple bias analysis

Following the works of Schlesselman (1978) and Bross (1966), we detail the method for exploring the effect of the unmeasured confounder. Given the assumptions about the distribution of the confounder in the population and the effect of the confounder on the outcome in the absence of exposure, the link between the observed risk ratios and the risk ratio adjusted for an unmeasured confounder is described as

$$RR_{adj} = RR_{obs}\left[\frac{RR_{UY}p_0 + (1 - p_0)}{RR_{UY}p_1 + (1 - p_1)}\right] \quad (1)$$

where RR_{adj} is the adjusted risk ratio, RR_{obs} is the observed risk ratio computed without adjusting for the unmeasured confounder, is the risk ratio associating the unmeasured confounder with the outcome, p_1 and p_0 are prevalence of the unmeasured confounder in the exposed and the unexposed groups. Mathematically, $p_1=P(U=1|X=1)$, $p_0=P(U=1|X=0)$. The RR_{obs} can be estimated from a generalised linear model or even a simple cross tabulation. This method is selected, to illustrate that one does not require special software to carry out bias analysis. Using hypothetical (e.g., collected from systematic reviews) estimates of the three parameters (RR_{UY}, p_0, p_1), one can calculate the association between the exposure and the disease after accounting for the unmeasured confounder. In the above formula, we are also making an assumption that the effect of on does not differ within the levels of exposure/treatment.

Alternatively, if one is aware of the joint distribution of exposure/treatment, outcome and the unmeasured confounder (all measured as dichotomous variables), then the data could be arranged as shown in Table 1, and the adjusted risk ratio or risk difference can be computed as detailed below:

Table 1: Association between exposure (X) and outcome (Y) stratified by the unmeasured confounder (U).

	Total		$U = 1$		$U = 0$	
Outcome	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$
$Y = 1$	a	b	A_1	B_1	A_0	B_0
$Y = 0$	c	d	C_1	D_1	C_0	D_0
	n_1	n_2	N_{11}	N_{01}	N_{10}	N_{00}

With the assumptions about the bias parameters, the information in the cells for the middle and right columns of Table 1 can be calculated using the information from the first two columns. To get the information on n_1 and n_2, and cells (*a, b, c, d*) one can perform a simple cross-tabulation using observed data. Then use the hypothetical estimates of the prevalence of the confounder among the exposed (p_1) and prevalence of the confounder among the unmeasured (p_0) to estimate

$$N_{11} = n_1 * p_1 \text{ and } N_{10} = n_1 - N_{11}$$

and

$$N_{01} = n_2 * p_0 \text{ and } N_{00} = n_2 - N_{01}.$$

Once the marginal totals (N_{11}, N_{10}, N_{01}, N_{00}) are computed, then we can compute the information that is required to fill the cells (A_1, B_1, C_1, D_1, A_0, B_0, C_0, D_0). Before this we need to decide on the parameter of interest (RR/RD). Let's say we are interested in estimating risk ratio - defined as

$$RR_{UY} = \frac{\frac{B_1}{N_{01}}}{\frac{B_0}{N_{00}}}$$

but one can substitute B_0 with ($b - B_1$) and N_{00} with ($n_2 - N_{01}$) and rearrange the terms to compute B_1, which is estimated as

$$B_1 = \frac{RR_{UY}N_{01}b}{RR_{UY}N_{01} + n_2 - N_{01}}$$

In the above equation RR_{UY} is an assumed value, N_{01} is computed using above equations and n_2 and b are estimated from data. Using all of this information now one can compute B_1.

Similarly, the value of A_1 can be computed as

$$A_1 = \frac{RR_{UY}N_{11}a}{RR_{UY}N_{11} + n_1 - N_{11}}$$

Once the cell information (A_1 and B_1) is estimated, then the adjusted risk ratio can be computed using Equation 1. Information on A_1, A_2, B_1, B_2, D_1 and D_2 are computed to complete the table. These values will not be used here. However, if one is interested in computing the odds ratio then these are required.

Risk Difference

When the parameter of interest is RD, the method for bias analysis (simple version) is detailed below. Similar to the risk ratio, we require information on the prevalence of the confounder in both the exposed and unexposed groups. One also needs to know the risk difference associating the confounder with the outcome (RD_{UY}). Computation of N_{11} and N_{10} and N_{01} and N_{00} is similar to how they were described in the risk ratio, but determination of the values A_1, B_1, A_2 and B_2 is done differently.

Theoretically, the risk difference is computed as (Lash, Fox and Fink, 2011; Lash and Fink 2003);

$$RD_{UY} = \frac{B_1}{N_{01}} - \frac{B_0}{N_{00}}$$

However we know from Table 1 that $B_0 = b - B_1$ and $N_{00} = n_2 - N_{01}$ now substituting these values in the above equation and rearranging them one can compute the required value of B_1 as

$$B_1 = \frac{RD_{UY}n_2(n_2 - N_{01}) + bN_{01}}{n_2}$$

Similarly, the value for can be computed as

$$A_1 = \frac{RD_{UY}n_1(n_1 - N_{11}) + aN_{11}}{n_1}$$

Now using these newly computed values of the cells, the adjusted risk difference can be computed as

$$RD_{adj} = RD_{obs} + (RD_{UY})(p_0 - p_1)$$

Example

Using hypothetical data, and the above method I demonstrate how one could estimate the adjusted risk ratio and risk difference for unmeasured confounding. I use a different example that considers the individual level aspects.

Suppose the interest is in knowing the risk of sugar consumption on tooth decay. Using collected information we estimate the risk ratio between sugar consumption (SC) and tooth decay (TD), but want to adjust this risk for an unmeasured confounder, e.g., oral health literacy (OHL). The user-supplied values for bias parameters are summarised in Table 2. Using this new information let's see how we can adjust the observed effect of SC on TD.

Table 2: Bias parameters for simple bias analysis of the association between sugar consumption (X) and tooth decay (Y) stratified by an unmeasured confounder U (oral health literacy).

Bias Parameter	*Description*	*User supplied values to the bias parameter*
RR_{UY}	Association between good oral health literacy and dental caries	1.5
p_1	Prevalence of good oral health literacy among people with tooth decay	0.35
p_0	Prevalence of good oral health literacy among people with no tooth decay	0.75

The left side of Table 3 shows the observed data associating SC (X) on TD (Y).

Table 3: Hypothetical Data on the relationship between sugar consumption and tooth decay stratified by oral health literacy (unmeasured confounder). Crude (left two columns) and estimated data (remaining four columns).

	Total		*U* = 1		*U* = 0	
Outcome	*X* = 1	*X* = 0	*X* = 1	*X* = 0	*X* = 1	*X* = 0
Y = 1	110	200	49.15	163.64	60.85	36.36
Y = 0	130	600	34.85	436.36	95.15	163.64
	240	800	84	600	156	200

The crude data and assumptions about the bias parameters, in conjunction with the equations shown in previous section, provide a solution to allow stratification by oral health literacy. The crude risk ratio can be computed from Table 3 as 1.84 ((110/240)/(200/800)). To get the adjusted estimates we start with solving for N_{11} and N_{01} using the prevalence values stated in Table 2.

$$N_{11} = n_1 * p_1 = 240 * 0.35 = 84$$
$$N_{01} = n_2 * p_0 = 800 * 0.75 = 600$$

Then for N_{10} and N_{00}

$$N_{10} = 240 - 84 = 156$$
$$N_{00} = 800 - 600 = 200$$

The relation between oral health literacy and tooth decay can be used to solve the cell values of the stratified Table 2

$$A_1 = \frac{RR_{UY}N_{11} * a}{RR_{UY}N_{11} + n_1 - N_{11}} = \frac{1.5 * 84 * 110}{1.5 * 84 + 240 - 84} = 49.15$$

And

$$B_1 = \frac{RR_{UY}N_{01}b}{RR_{UY}N_{01} + n_2 - N_{01}} = \frac{1.5 * 600 * 200}{1.5 * 600 + 800 - 600} = 163.64$$

The estimated, stratified data, allows calculation of the adjusted association between SC and TD for confounding by OHL. After adjusting for OHL, the standardised relative risk can be estimated as

$$RR_{adj} = \frac{110}{84 * \left(\frac{163.64}{600}\right) + \left(\frac{36.36}{200}\right) * 156} = 2.15$$

Let's use the formula and compute the adjusted risk ratio

$$RR_{adj} = RR_{obs}\left[\frac{RR_{UY}p_0 + (1 - p_0)}{RR_{UY}p_1 + (1 - p_1)}\right] = 1.84 * \left[\frac{1.5 * 0.75 + (1 - 0.75)}{1.5 * 0.35 + (1 - 0.35)}\right] = 2.15$$

We can now see the impact of the bias caused by unmeasured confounding. The adjusted risk ratio is higher than the observed risk ratio. In this example, the risk was underestimated when the unmeasured confounding was not considered. One can argue that the estimates from bias analysis are as good as the values assigned to the bias parameters. While true, the point to emphasise here is that presenting the results from a quantitative bias analysis is an improvement over intuitive estimates of the impact of unmeasured confounding because the assumptions are explicit and the impacts given those assumptions can be quantified. Additionally, bias analysis allows evaluation of the plausibility of competing evaluations for the observed associations. As seen from the above example, bias analysis can provide justification for collection of new information if the gain in information due to unmeasured confounding is expected to be marginal. However, this must not lead to conclusions such as absence of bias implies no new information is required. If one does not want to limit to a single value of the bias parameters, then one can also define multiple values for any of the three parameters. For example, computing the adjusted risk ratio for a combination of values of p_0 and p_1 fixing RR_{UY}, which can also be done using the above method (see R code in web appendix). The results then can be presented as shown in Figure 2.

While the method described here is simple and allows to explore the impact of unmeasured confounder, it is limited by the accuracy of the values assigned to parameters and the knowledge of unmeasured/unknown confounders. Moreover the bias analysis presented above was carried out under the assumption that there was only a single unmeasured confounder. In reality, there can be several confounders; in such cases this method is not applicable and one must use more advanced methods such as probabilistic bias analysis.

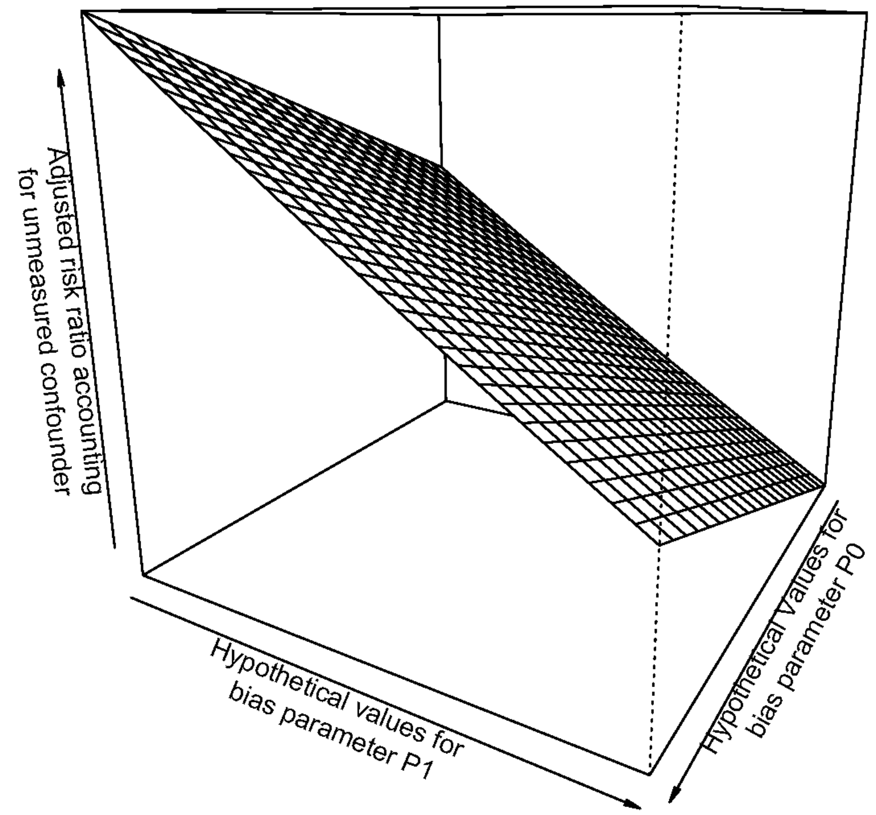

Figure 2: 3D surface plot of adjusted risk ratios for a sequence of bias parameters p1 and p0.

Risk Difference

Here I demonstrate the conduct of bias analysis for risk difference. We still use the information presented Table 3 for the first two columns.

Table 4: Hypothetical data on the relationship between sugar consumption and tooth decay stratified by oral health literacy (Unmeasured confounder). Crude (first two columns) and estimated data (remaining four columns).

	Total		$U = 1$		$U = 0$	
Outcome	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$
$Y = 1$	110	200	61.9	180	48.1	20
$Y = 0$	130	600	22.1	420	107.9	180
	240	800	84	600	156	200

Values of N_{11}, N_{01}, N_{10} and N_{00} are computed as described above for risk ratio. However for computing the values of A_1 and B_1 we need the risk difference values of bias parameter and the values of prevalence. These are assumed to be RD_{UY}= 0.15, p_0= 0.75 and p_1= 0.35.

$$B_1 = \frac{RD_{UY} n_2 (n_2 - N_{01}) + bN_{01}}{n_2} = \frac{0.15 * 800 * (800 - 600) + 200 * 600}{800} = 180$$

Similarly the value for can be computed as

$$A_1 = \frac{RD_{UY} n_1 (n_1 - N_{11}) + aN_{11}}{n_1} = \frac{0.15 * 240 * (240 - 84) + 110 * 84}{240} = 61.9$$

Now using these estimated values of the cells the adjusted risk difference can be computed as

$$RD_{adj} = RD_{obs} + (RD_{UY})(p_0 - p_1) = 0.21 + 0.15 * (0.75 - 0.35) = 0.27$$

Similar to the risk ratio, once again we notice that risk difference is underestimated when the unmeasured confounder (oral health literacy) is not accounted for. (See web supplement for the R code used in the computation https://universityofadelaide.box.com/s/7n3fbyaqkh1oevl2a0h0fvvdmuwd2nte).

Discussion

Usually, when analysing the results of an epidemiological study, the true data generating model is never known. We do not reliably know which variables are confounders of the association of interest, the form in which they should enter the model, or the time scale over which they act. Therefore, validity of an epidemiological study may be threatened by both residual and unmeasured confounding. As shown repeatedly, unmeasured confounding can be a serious problem (Greenland, 1996; Rosenbaum and Rubin, 1983, Brumback *et al*., 2004). Where residual confounding is defined as the distortion that remains after controlling for confounding in design and/or analysis of a study (e.g confounding due to measurement error). As demonstrated in this paper, even when one unmeasured confounder is omitted from analysis, it can lead to biased estimates. Confounding can be caused by variables that are associated with both outcome and exposure and are not on the causal pathway between exposure and outcome. Controlling for variables with these properties may remove bias; investigators must perfectly characterise their association with the exposure of interest. If such characterisation is not possible, then one must perform sensitivity analysis to assess whether unmeasured and residual confounding are likely problems. Even though we have illustrated how the sensitivity analysis is performed for single confounders, the possibility of the presence of several unmeasured confounders should not be ruled out. Many authors have shown that unmeasured confounders can have a cumulative effect (Greenland, 1996; Lash, 2003; Brumback *et al.*, 2004). Therefore it may not be enough to state that a single unmeasured confounder would need an implausibly large relative risk to remove the observed confounding. Here we have used hypothetical examples to illustrate the importance of unmeasured confounding. While analytic results are always desirable, they are not always possible to attain (Van der Weele and Arah, 2011). Moreover, from a logistic point of view, epidemiological data collection and analysis often require substantial resources including financial resources to collect data, hence these resources must be spent with great care especially if the gain in information is expected to marginal.

Acknowledgements

I would like to thank Dr. Janet Grant for helping me with the language correction, and Drs. Manasi Mittinty and Dandara Haag for providing meaningful discussions on the previous versions of the manuscript.

References

Bross, I.D. (1966): Spurious effects from an extraneous variable. *Journal of Chronic Diseases* **19**, 637-647.

Brumback, B.A., Hernán, M.A., Haneuse, S.J. and Robins, J.M. (2004): Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine* **23**, 749-767.

Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B. and Wynder, E.L. (1959): Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute* **22**, 173-203.

Greenland, S. (1996): Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology* **25**, 1107-1116.

Hernán M.A. and Robins J.M. (2020): *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.

Lash, T.L., Fox, M.P., MacLehose, R.F., Maldonado, G., McCandless, L.C. and Greenland, S. (2014): Good practices for quantitative bias analysis. *International Journal of Epidemiology* **43**, 1969-1985.

Lash, T.L., Fox, M.P. and Fink, A.K. (2011): *Applying quantitative bias analysis to epidemiologic data.* Springer Science & Business Media.

Lash, T.L. and Fink, A.K. (2003): Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology* **14**, 451-458.

Rosenbaum, P.R. and Rubin, D.B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)* **45**, 212-218.

Schlesselman, J.J. (1978): Assessing effects of confounding variables. *American Journal of Epidemiology* **108**, 3-8.

Sterne, J.A., Hernán, M.A., Reeves, B.C., Savović, J., Berkman, N.D., Viswanathan, M., Henry, D., Altman, D.G., Ansari, M.T., Boutron, I. and Carpenter, J.R. (2016): ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* **355**, i4919.

Van der Weele, T.J. and Arah, O.A. (2011): Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* **22**, 42-52.