Bias Analysis

EPIB 705

Sam Harper

2023-03-22

Overview

Why Bias Analysis?

Deterministic Bias Analysis

Unmeasured confounding

Misclassification

Selection bias

Multidimensional Bias Analysis

Record-level Implementation

Summary

Overview

Why Bias Analysis?

Deterministic Bias Analysis

Unmeasured confounding

Misclassification

Selection bias

Multidimensional Bias Analysis

Record-level Implementation

Summary

Introduction

How do misclassification, selection bias, and unmeasured confounding create bias in parameter estimation?

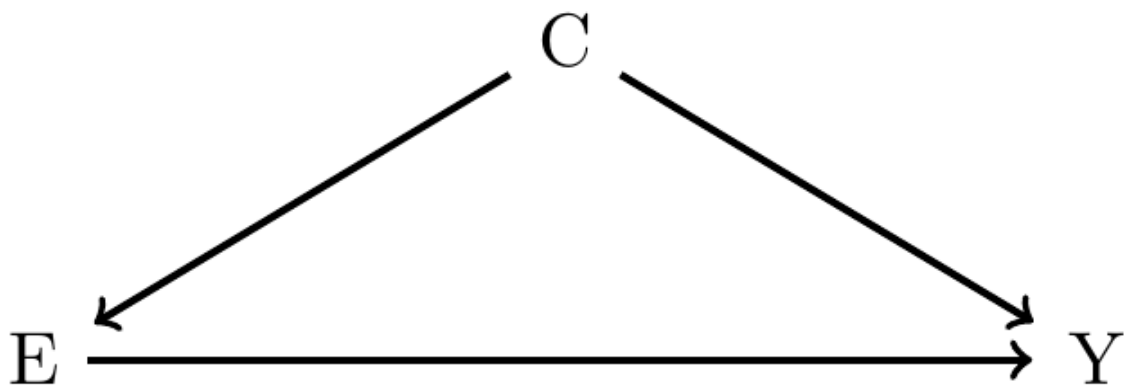
> By assuming that all errors are random and that any modeling assumptions (such as homogeneity) are correct, all uncertainty about the effect of errors on estimates is subsumed within conventional standard deviations for the estimates (standard errors), such as those given in earlier chapters (which assume no measurement error), and any discrepancy between an observed association and the target effect may be attributed to chance alone[1]

- Bias analysis is an attempt to quantify the potential for bias, and reduce the likelihood of mistakenly attributing effects to exposure rather than systematic error.

1. See Rothman et al. (2008), p.362.

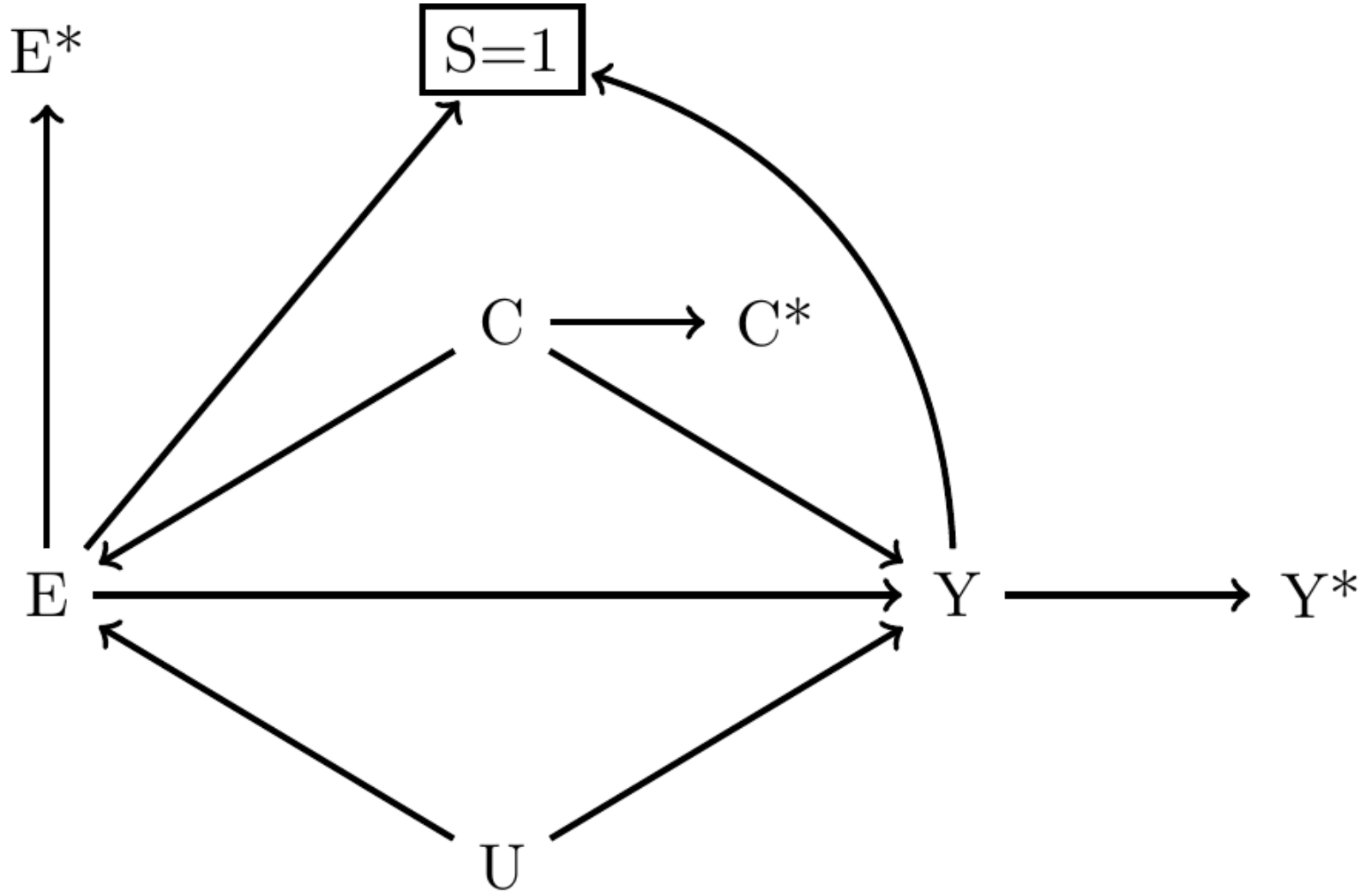
What are we concerned about?

In theory.



What are we concerned about?

In practice.



Potential consequences

- “Bias analysis requires educated guesses about the likely sizes of systematic errors”[1]

- The difficulty is that it is challenging to do so quantitatively, thus investigators often rely on qualitative judgments about the likelihood that their estimates are biased.

- A basic problem is that the results of observational studies are likely to be sensitive to choices made by the analyst.

1. Rothman, Greenland, Lash, *Modern Epidemiology* 3rd ed, p,347.

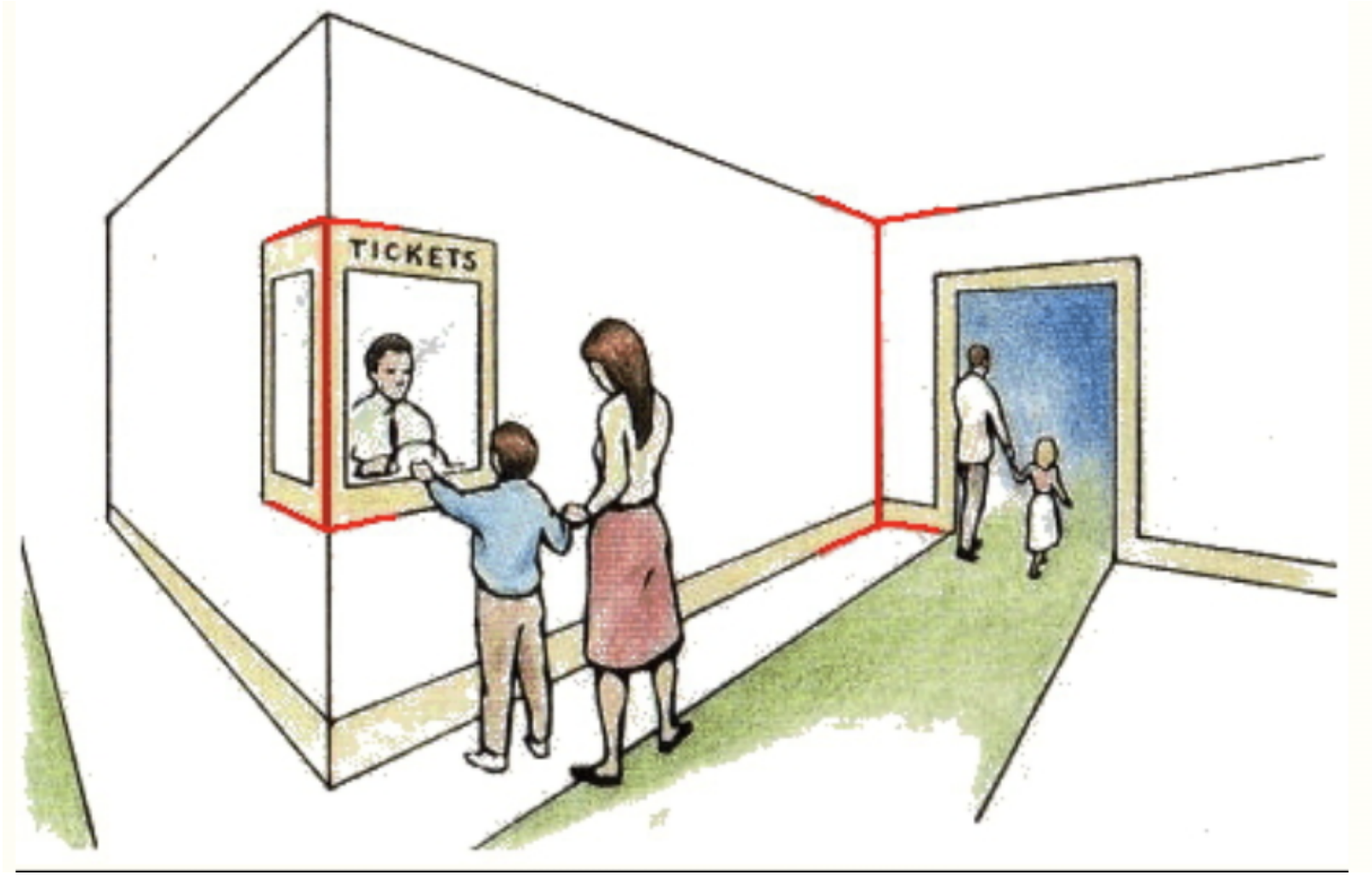
Example of typical qualitative analysis

Study[1] assessing the association of vitamin D on age-related macular degeneration (AMD):

> Several potential limitations of the present investigation must be considered in drawing conclusions from the results. In particular, AMD was ascertained only in one eye, resulting in a possible underestimation of AMD cases; however, studies have shown that AMD development is typically symmetric. Further, AMD was identified using nonmydriatic fundus photography without dilating the pupils, which may have led to potential misclassification of cases. In estimating milk and fish intake, the food frequency questionnaire used in this study was not validated and the measurement error was unknown. The serum 25-hydroxyvitamin D values would reflect sun exposure and food intake over recent weeks, rather than years, which would have enhanced random measurement error. Therefore, **associations reported are likely to be biased toward the null.**

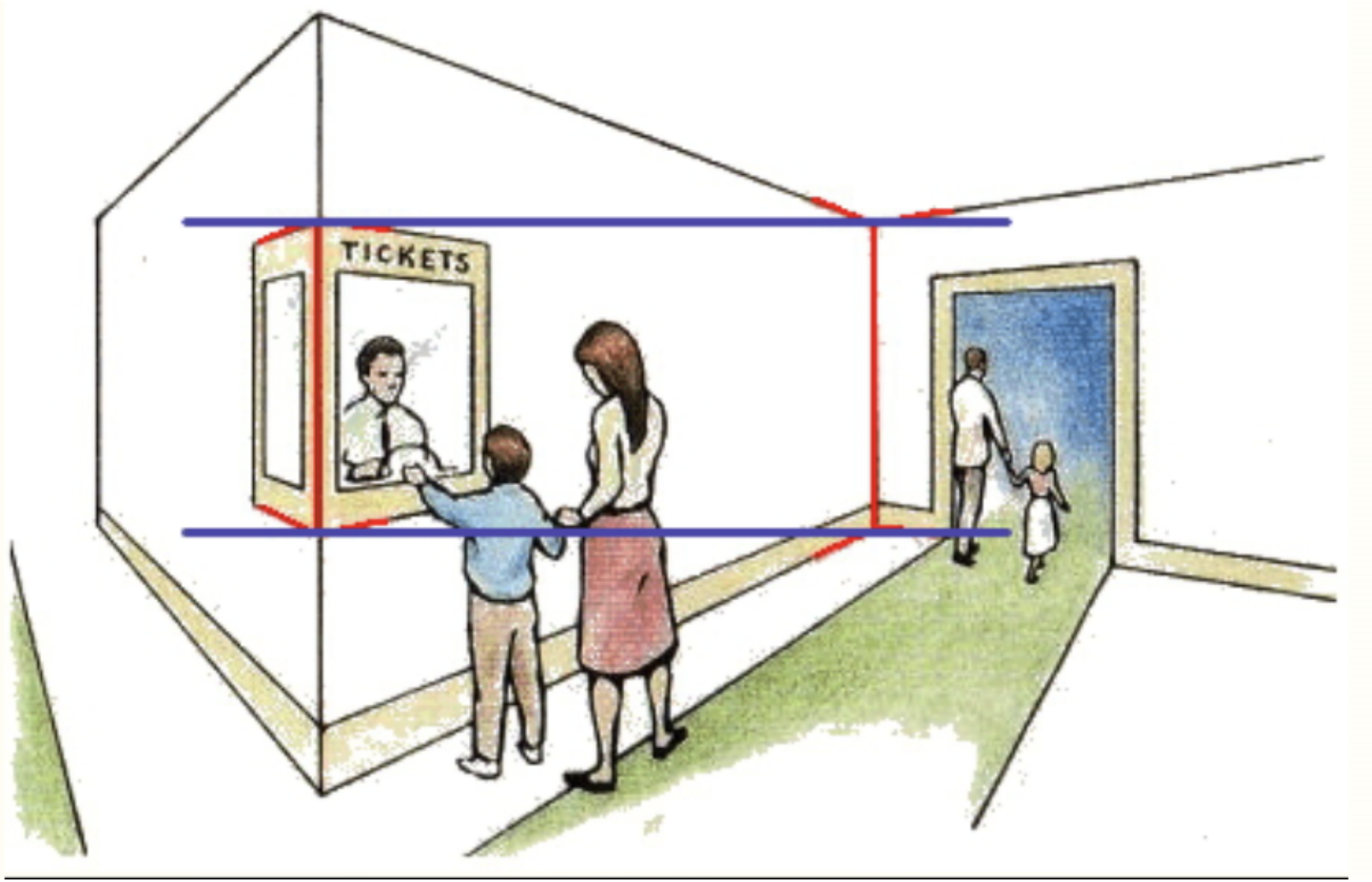
1. Parekh et al. (2007)

Which red line is longer?[1]



1. https://michaelbach.de/ot/sze-muelue/

Intuitions are difficult to overcome.[1]



1. https://michaelbach.de/ot/sze-muelue/

Overview

Why Bias Analysis?

Deterministic Bias Analysis

Unmeasured confounding

Measurement error

Selection bias

Multidimensional Bias Analysis

Record-level Implementation

Summary

Rationale for bias analysis for unmeasured confounding

When should you conduct a bias analysis for unmeasured confounding?

1. An important (and well known) confounder was not measured (e.g., too expensive to collect; reliance on secondary data)
2. Quantifying the impact of an unknown confounder (e.g., early study of an association; unmeasured confounding seems likely)

. . .

- Goals:
  - We want to provide an estimate of the association corrected for the unmeasured confounder (Z).
  - Answers the question of what the association would have been **had we controlled for** Z.

Where did this come from?

- Early ideas formulated in the context of controversy surrounding the link between smoking and lung cancer[1]

> If a causal agent, *A*, with no causal effect upon the risk of a disease, nevertheless, because of a positive correlation with some other causal agent, *B*, shows an apparent risk, *r*, for those exposed to *A*, relative to those not so exposed, then the prevalence of *B*, among those exposed to *A*, relative to the prevalence among those not so exposed, must be greater than *r*.

Jerome Cornfield, 1974

1. Bross (1954) is credited with implementing the first quantitative bias analyis, but Cornfield et al. (1959) also played an important role.

Unmeasured Confounding

- Suppose we want to assess the effect of occupational exposure to resins (X) on lung cancer risk (D) using a case-control design.[1]

Disease Status	X=1	X=0	Total
Cases (D=1)	45	94	139
Controls (D=0)	257	945	1202
Total	302	1039	1341

Crude OR:

$$OR_{DX+} = \frac{(45 \times 945)}{(94 \times 257)} = 1.76$$

- Likely confounded by smoking (Z), but we did not measure it.
- How can we quantify failing to adjust for Z?

1. Longstanding example used in *Modern Epidemiology* from Greenland et al. (1994)

Data Layout for Dichotomous Unmeasured Confounder

	Z=1 Smokers			Z=0 Non-smokers			Total		
2-4	X=1	X=0	Total	X=1	X=0	Total	X=1	X=0	Total
D=1	A_{11}	A_{01}	M_{11}	$A_{1+}-A_{11}$	$A_{0+}-A_{01}$	$M_{A+}-M_{11}$	A_{1+}	A_{0+}	M_{A+}
D=0	B_{11}	B_{01}	M_{01}	$B_{1+}-B_{11}$	$B_{0+}-B_{01}$	$M_{B+}-M_{01}$	B_{1+}	B_{0+}	M_{B+}
Total	N_{11}	N_{01}	N_{+1}	$N_{1+}-N_{11}$	$N_{0+}-N_{01}$	$N_{++}-N_{+1}$	N_{1+}	N_{0+}	N_{++}

- The crude odds ratio is $OR_{DX+} = A_{1+}B_{0+}/A_{0+}B_{1+}$, but to adjust for smoking we need smoking stratum-specific ORs.
- We can write the exposure-disease OR **within strata** of Z (OR_{DXZ}) as:

$$OR_{DX1} = \frac{A_{11}B_{01}}{A_{01}B_{11}} \quad \text{and} \quad OR_{DX0} = \frac{(A_{1+} - A_{11})(B_{0+} - B_{01})}{(A_{0+} - A_{01})(B_{1+} - B_{11})}$$

Data Layout for Dichotomous Unmeasured Confounder

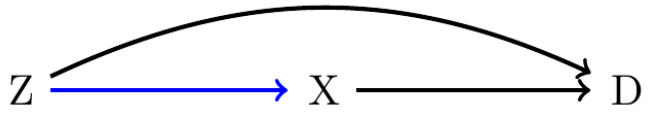
We already know the marginal totals, so only 4 quantities to estimate.

	Z=1 Smokers			**Z=0 Non-smokers**			**Total**		
2-4	X=1	X=0	Total	X=1	X=0	Total	X=1	X=0	Total
D=1	$\mathbf{A_{11}}$	$\mathbf{A_{01}}$	M_{11}	$45-A_{11}$	$94-A_{01}$	$139-M_{11}$	45	94	139
D=0	$\mathbf{B_{11}}$	$\mathbf{B_{01}}$	M_{01}	$257-B_{11}$	$945-B_{01}$	$1202-M_{01}$	257	945	1202
Total	N_{11}	N_{01}	N_{+1}	$302-N_{11}$	$1039-N_{01}$	$1341-N_{+1}$	302	1039	1341

- These values may be generated by specifying values for the prevalence of smoking in each exposure group and the association between smoking and disease.
- The (unknown) smoking prevalence in exposure strata are:

$$P_{Z1} = N_{11}/N_{1+} \quad \text{and} \quad P_{Z0} = N_{01}/N_{0+}$$

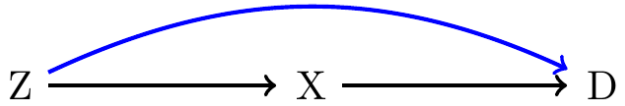
What do we need? (1) Distribution of Z by exposure



	Z=1 Smokers			Z=0 Non-smokers			Total		
2-4	X=1	X=0	Total	X=1	X=0	Total	X=1	X=0	Total
D=1	$\mathbf{A_{11}}$	$\mathbf{A_{01}}$	M_{11}	$45-A_{11}$	$94-A_{01}$	$139-M_{11}$	45	94	139
D=0	$\mathbf{B_{11}}$	$\mathbf{B_{01}}$	M_{01}	$257-B_{11}$	$945-B_{01}$	$1202-M_{01}$	257	945	1202
Total	N_{11}	N_{01}	N_{+1}	$302-N_{11}$	$1039-N_{01}$	$1341-N_{+1}$	302	1039	1341

- With plausibly valid estimates of P_Z, we can use this to estimate the number of controls, since $B_{11} = P_{Z1}B_{1+}$ and $B_{01} = P_{Z0}B_{0+}$.
- However, A_{11} and A_{01} are still unknown, but note that they can be estimated if we specify a plausible value for the association between smoking (Z) and disease (D).

What do we need? (2) Effect of Z on D



	Z=1 Smokers			Z=0 Non-smokers			Total		
2-4	X=1	X=0	Total	X=1	X=0	Total	X=1	X=0	Total
D=1	A_{11}	A_{01}	M_{11}	$A_{1+}-A_{11}$	$A_{0+}-A_{01}$	$M_{A+}-M_{11}$	A_{1+}	A_{0+}	M_{A+}
D=0	B_{11}	B_{01}	M_{01}	$B_{1+}-B_{11}$	$B_{0+}-B_{01}$	$M_{B+}-M_{01}$	B_{1+}	B_{0+}	M_{B+}
Total	N_{11}	N_{01}	N_{+1}	$N_{1+}-N_{11}$	$N_{0+}-N_{01}$	$N_{++}-N_{+1}$	N_{1+}	N_{0+}	N_{++}

- Based on the table above, the confounder-disease OR within strata of X (OR_{DZX}) can be calculated as:

$$OR_{DZ1} = \frac{A_{11}\left(B_{1+} - B_{11}\right)}{\left(A_{1+} - A_{11}\right)B_{11}}, \quad \text{and} \quad OR_{DZ0} = \frac{A_{01}\left(B_{0+} - B_{01}\right)}{\left(A_{0+} - A_{01}\right)B_{01}}$$

Estimating Z among cases

Recall our stratum-specific ORs:

$$OR_{DZ1} = \frac{A_{11}\left(B_{1+} - B_{11}\right)}{\left(A_{1+} - A_{11}\right)B_{11}} \quad \text{and} \quad OR_{DX0} = \frac{A_{01}\left(B_{0+} - B_{01}\right)}{\left(A_{0+} - A_{01}\right)B_{01}}$$

- If we can substitute reasonable values for these z-specific ORs, we can solve[1] the above equations to get the value of A_{11} and A_{01}:

$$A_{11} = OR_{DZ1}A_{1+}B_{11}/(OR_{DZ1}B_{11} + B_{1+} - B_{11})$$

$$A_{01} = OR_{DZ0}A_{0+}B_{01}/(OR_{DZ0}B_{01} + B_{0+} - B_{01})$$

1. See Rothman et al. (2008), p.350, eq. 19-1 and 19-2

Worked example

- Suppose we know from external studies that the prevalence of smoking is 70% among those occupationally exposed to resins and 50% among those unexposed in this population.

	Z=1 Smokers			**Z=0 Non-smokers**			**Total**		
2-4	X=1	X=0	Total	X=1	X=0	Total	X=1	X=0	Total
D=1	$\mathbf{A_{11}}$	$\mathbf{A_{01}}$	M_{11}	$45-A_{11}$	$94-A_{01}$	$139-M_{11}$	45	94	139
D=0	180	473	M_{01}	$257-B_{11}$	$945-B_{01}$	$1202-M_{01}$	257	945	1202
Total	N_{11}	N_{01}	N_{+1}	$302-N_{11}$	$1039-N_{01}$	$1341-N_{+1}$	302	1039	1341

- Our estimates of B_{11} and B_{01} are now:

$$B_{11} = P_{Z1}B_{1+} = (.7)(257) \approx 180$$
$$B_{01} = P_{Z0}B_{0+} = (.5)(945) \approx 473$$

	Z=1 Smokers			Z=0 Non-smokers			Total		
2-4	X=1	X=0	Total	X=1	X=0	Total	X=1	X=0	Total
D=1	41	77	M_{11}	$45-A_{11}$	$94-A_{01}$	$139-M_{11}$	45	94	139
D=0	180	473	M_{01}	$257-B_{11}$	$945-B_{01}$	$1202-M_{01}$	257	945	1202
Total	N_{11}	N_{01}	N_{+1}	$302-N_{11}$	$1039-N_{01}$	$1341-N_{+1}$	302	1039	1341

- Plugging in the estimates of B_{11} and B_{01} along with a plausible estimate of the confounder-disease association ($OR_{DZ1} = OR_{DZ0} = 5$), assumed[1] to be homogeneous across strata of resin exposure, now allow us to calculate A_{11} and A_{01}:

$$A_{11} = \frac{OR_{DZ1}A_{1+}B_{11}}{OR_{DZ1}B_{11}+B_{1+}-B_{11}} = \frac{5(45)(180)}{5(180)+257-180} \approx 41$$

$$A_{01} = \frac{OR_{DZ0}A_{0+}B_{01}}{OR_{DZ0}B_{01}+B_{0+}-B_{01}} = \frac{5(94)(473)}{5(473)+945-473} \approx 77$$

1. Not necessary to assume homogeneity if there is evidence to the contrary.

We can now fill out the table and calculate an OR standardized for smoking:

	Z=1 Smokers			**Z=0 Non-smokers**			**Total**		
2-4	X=1	X=0	Total	X=1	X=0	Total	X=1	X=0	Total
D=1	41	78	120	4	16	19	45	94	139
D=0	180	473	652	77	473	550	257	945	1202
Total*	221	551	772	81	488	569	302	1039	1341

*Note: Row and column totals in Z strata may not sum because of rounding.

- If we standardize to the exposed population, we get:

$$OR_{DXZ(E)} = \frac{\sum_z B_{1z}(A_{1z}/B_{1z})}{\sum_z B_{1z}(A_{0z}/B_{0z})} = \frac{180(41/180) + 77(4/77)}{180(78/473) + 77(16/473)} = 1.39$$

Simplifying formula

- Bias relative to the crude estimate:

$$Bias(OR) = \frac{OR_{DX+}}{OR_{DXZ(E)}} = \frac{1.76}{1.39} = 1.27$$

- Arah et al.[1] also give an alternative estimator using only 1) prevalence of Z in each exposure stratum and 2) association between Z and D

$$Bias(OR) = \frac{OR_{DX+}}{OR_{DXZ(E)}} = \frac{OR_{DZ0}P_{11} + 1 - P_{11}}{OR_{DZ0}P_{10} + 1 - P_{10}} = \frac{5(.7) + 1 - .7}{5(.5) + 1 - .5} = 1.27$$

where $P_{11} = P(Z = 1|X = 1)$, $P_{10} = P(Z = 1|X = 0)$, and OR_{DZ0} is the $Z \to D$ association in the unexposed.

1. See Arah et al. (2008)

Extensions to other effect measures

- The logic for adjusting the OR above applies equally to other effect measures (RD, RR), with minor alterations.
- For both RD and RR you need to specify the difference in the prevalence of Z by exposure
- For RR need to specify the association between Z and D on the RR scale.

- In our example above, we specified $OR = 5$ for $Z \rightarrow D$ association, roughly $RR = 4.32$.
- The bias in the crude RR is therefore:

$$Bias(RR) = \frac{RR_{DX+}}{RR_{DXZ(E)}} = \frac{RR_{DZ0}P_{11} + 1 - P_{11}}{RR_{DZ0}P_{10} + 1 - P_{10}} = \frac{4.3(.7) + 1 - .7}{4.3(.5) + 1 - .5} = 1.25$$

What about the risk difference?[1]

- Still need the prevalence of the unmeasured confounder by exposure, but now we need the risk difference RD_{DZ} for the confounder-disease association.

	Z=1 Smokers			Z=0 Non-smokers			Total		
2-4	X=1	X=0	Total	X=1	X=0	Total	X=1	X=0	Total
D=1	$\mathbf{A_{11}}$	$\mathbf{A_{01}}$	M_{11}	$45-A_{11}$	$94-A_{01}$	$139-M_{11}$	45	94	139
D=0	$\mathbf{B_{11}}$	$\mathbf{B_{01}}$	M_{01}	$257-B_{11}$	$945-B_{01}$	$1202-M_{01}$	257	945	1202
Total	N_{11}	N_{01}	N_{+1}	$302-N_{11}$	$1039-N_{01}$	$1341-N_{+1}$	302	1039	1341

- The crude association is $RD_{DX+} = (45/302) - (94/1039) = 0.06$
- Difference for Z by exposure (70% $X = 1$, 50% $X = 0$)
- Choose RD for $Z \rightarrow D$ among the unexposed as $RD_{DZ0} = 0.10$.

1. These data are derived from a case-control study, so this is just for illustration.

Revised table for RD:

	Z=1 Smokers			Z=0 Non-smokers			Total		
2-4	X=1	X=0	Total	X=1	X=0	Total	X=1	X=0	Total
D=1	38	73	111	7	21	28	45	94	139
D=0	174	447	620	83	499	582	257	945	1202
Total*	211	520	731	91	520	610	302	1039	1341

*Note: Row and column totals in Z strata may not sum because of rounding.

- Standardize to exposed, we get:

$$RD_{DXZ(E)} = \sum_z w_z RD_{DXZ} / \sum_z w_z = 0.04$$

- Bias is $RD_{DX+} - RD_{DXZ(E)} = 0.06 - 0.04 = 0.02$
- Arah et al. show how to calculate simply as:
  $Bias = RD_{DX+} - RD_{DXZ(E)} = RD_{DZ0}\left(P_{Z1} - P_{Z0}\right) = 0.10\left(0.7 - 0.5\right) = 0.02$

General formulas based on counterfactual notation

- What to do for more complex, regression-based models?[1]
- VanderWeele and Arah (2011) provided some general formulas.
- The average causal effect among those exposed, which **would** be adjusted for both measured Xs **and** unmeasured U is:

$$E\left(Y_{a1}|a_1\right) - E\left(Y_{a0}|a_1\right)$$

$$= \sum_x \sum_u \{E(Y|a_1, x, u) - E(Y|a_0, x, u)\} P(u|x, a_1) P(x|a_1)$$

1. Far more likely what you'll be using in practice. See Vanderweele and Arah (2011). More on this later.

General formulas for unmeasured confounding

- But, without adjustment for U we get an X-adjusted effect of:

$$\sum_x \{E(Y|a_1, x) - E(Y|a_0, x)\}P(x|a_1)$$

- So, the difference between these two estimates, d_{a1}, is

$$d_{a1} = \sum_x \{E(Y|a_1, x) - E(Y|a_0, x)\}P(x|a_1) - \{E\left(Y_{a1}|a_1\right) - E\left(Y_{a0}|a_1\right)\}$$

General formulas for unmeasured confounding

- The bias is (still) a function of the prevalence of the unknown confounder and its effect on the outcome (reference level=u'):

$$d_{a1} = \sum_x \sum_u \left\{ E(Y|a_0, x, u) - E(Y|a_0, x, u') \right\} \left\{ P(u|a_1, x) - P\left(u|a_0, x\right) \right\} P\left(x|a_1\right)$$

We thus need to specify:

- The $U \rightarrow Y$ relationship among the unexposed:

$$E(Y|a_0, x, u) - E(Y|a_0, x, u')$$

- Distribution of U among exposed and unexposed *within strata of X*:

$$\{P(u|a_1, x) - P(u|a_0, x)\} P(x|a_1)$$

Complications for the general formula

- These formulas require a lot of knowledge.
  - Association between U and Y at each level of measured confounders X.
  - Association between U and A at each level of measured confounders X.
- How many measured confounders do you have? 2? 10?
- Even for 2 age groups and gender, for example, you would need to specify the $U \rightarrow Y$ and $U \rightarrow A$ relations for:
  - younger women, older women
  - younger men, older men
- Is this information available? Can you make educated guesses?

General formulas for unmeasured confounding

Simplifying assumptions could be useful. Assuming no heterogeneity across X, we could posit:

- constant prevalence difference between exposed and unexposed:

$$\delta = P(U = 1|a_1, x) - P(U = 1|a_0, x)$$

- constant RD for exposure to U across strata of exposure and X:

$$\gamma = E(Y|a, x, U = 1) - E(Y|a, x, U = 0)$$

Then the extent of bias is the product of these two terms:

$$d_{a1} = \gamma\delta$$

External adjustment for unmeasured confounding

Assumptions:[1]

- Differences in the distribution of U by exposure similar across all strata of measured Xs.
- Effect of U on Y similar across all strata of measured Xs.

Example Post-Hoc Bias Analysis	Rosenbaum & Rubin (1983)	
	Enough to nullify	More plausible
Adjusted estimate (0.67 – 0.36)	0.31 (0.17, 0.45)	0.31 (0.17, 0.45)
$\delta = P(U\|a_1) - P(U\|a_0)$ all strata	0.6	0.3
$\gamma = E(Y\|U = 1) - E(Y\|U = 0)$ all strata	0.517	0.5217
Bias$(\gamma\delta)$	0.310	0.155
Bias-corrected effect	0.0 (-0.14, 0.14)	0.16 (0.01, 0.30)

1. See Rosenbaum and Rubin (1983) for more details. Similar in some ways to the 'e-value' concept from VanderWeele and Ding (2017).

Break! ☕

07:00

Overview

Why Bias Analysis?

Deterministic Bias Analysis

Unmeasured confounding

Misclassification

Selection bias

Multidimensional Bias Analysis

Record-level Implementation

Summary

Epidemiology Faces Its Limits

The search for subtle links between diet, lifestyle, or environmental factors and disease is an unending source of fear—but often yields little certainty

The news about health risks comes thick and fast these days, and it seems almost constitutionally contradictory. In January of last year, for instance, a Swedish study found a significant association between residential radon exposure and lung cancer. A Canadian study did not. Three months later, it was pesticide residues. The *Journal of the National Cancer Institute* published a study in April reporting—contrary to previous, less powerful studies—that the presence of DDT metabolites in the bloodstream seemed to have no effect

ROBERT VISSER/GREENPEACE

Anxiety epidemic. Protesting risks that may—or may not—be real.

on the press for its reporting of epidemiology, and even on the public "for its unrealistic expectations" of what modern medical research can do for their health. But many epidemiologists interviewed by *Science* say the problem also lies with the very nature of epidemiologic studies—in particular those that try to isolate causes of noninfectious disease, known variously as "observational" or "risk-factor" or "environmental" epidemiology.

The predicament of these studies is a simple one: Over the past 50 years, epidemiologists have succeeded

Rothman, editor of the journal *Epidemiology*: "We're pushing the edge of what can be done with epidemiology."

With epidemiology stretched to its limits or beyond, says Dimitrios Trichopoulos, head of the epidemiology department at the Harvard School of Public Health, studies will inevitably generate false positive and false negative results "with disturbing frequency." Most epidemiologists are aware of the problem, he adds, "and tend to avoid causal inferences on the basis of isolated studies or even groups of studies in the absence of compelling biomedical evidence. However, exceptions do occur, and their frequency appears to be increasing." As Trichopoulos explains, "Objectively the problems are not more than they used to be, but the pressure is greater on the profession, and the number who practice it is greater."

As a result, journals today are full of stud-

Epidemiology Monitor:[1]

> That was one of the criticisms of your article. Epidemiologists said it is unbalanced and that you were only talking about our warts. What about our victories?

Gary Taubes:

> Well, what I am saying is the warts are huge. The victories are few, and at this point, a whole field may be on the verge of propagating pathological science, which means they cannot get good enough resolution to identify the effects they're studying. Epidemiologists may be seeing and reporting that there are canals on Mars because they're looking at Mars through Galileo's telescope. And that's the nature of the field and all the statistical wizardry in the world isn't going to change that because the experimental subjects are messy and the artifacts and biases found are so huge and the signals are small. Epidemiologists have to be willing to confront that. That's the problem.

1. Interview in Epidemiology Monitor of Gary Taubes, who published "Epidemiology Faces its Limits" in *Science* 1995.

SCIENCE ADVANCES | RESEARCH ARTICLE

EPIDEMIOLOGY

Epidemiology beyond its limits

Lauren E. McCullough[1]*†, Maret L. Maliniak[1]†, Avnika B. Amin[1], Julia M. Baker[1], Davit Baliashvili[1], Julie Barberio[1], Chloe M. Barrera[1], Carolyn A. Brown[2], Lindsay J. Collin[3], Alexa A. Freedman[4], David C. Gibbs[1], Maryam B. Haddad[1], Eric W. Hall[5], Sarah Hamid[1], Kristin R. V. Harrington[1], Aaron M. Holleman[1], John A. Kaufman[1], Mohammed A. Khan[1], Katie Labgold[1], Veronica C. Lee[1], Amyn A. Malik[6], Laura M. Mann[1], Kristin J. Marks[1], Kristin N. Nelson[1], Zerleen S. Quader[1], Katherine Ross-Driscoll[7], Supriya Sarkar[8], Monica P. Shah[1], Iris Y. Shao[1], Jonathan P. Smith[9], Kaitlyn K. Stanhope[10], Marisol Valenzuela-Lara[1], Miriam E. Van Dyke[1], Kartavya J. Vyas[1], Timothy L. Lash[1]

In 1995, journalist Gary Taubes published an article in *Science* titled "Epidemiology faces its limits," which questioned the utility of nonrandomized epidemiologic research and has since been cited more than 1000 times. He highlighted numerous examples of research topics he viewed as having questionable merit. Studies have since accumulated for these associations. We systematically evaluated current evidence of 53 example associations discussed in the article. Approximately one-quarter of those presented as doubtful are now widely viewed as causal based on current evaluations of the public health consensus. They include associations between alcohol consumption and breast cancer, residential radon exposure and lung cancer, and the use of tanning devices and melanoma. This history should inform current debates about the reproducibility of epidemiologic research results.

1

1. See McCullough et al. (2022). I'll leave to to you as to whether you find this a convincing rebuttal.

You may recall[1]

Two Antibody Studies Say Coronavirus Infections Are More Common Than We Think. Scientists Are Mad.

In California, two of the nation's first big antibody surveys estimated that the true number of coronavirus infections is significantly higher than believed. But scientists are skeptical.

Stephanie M. Lee
BuzzFeed News Reporter

Posted on April 22, 2020 at 6:35 pm

Right-wing and libertarian sites immediately seized on the findings, arguing that the economic shutdown has not been worth the public health gains.

Most experts agree there are far more coronavirus infections in the world than are being counted. But almost as instantly as the California numbers were released, critics called out what they saw as significant problems with, or at least big questions about, how the scientists had arrived at them. Chief among their concerns was the accuracy of the test underpinning both studies, and whether the scientists had fully accounted for the number of false positives it might generate.

1. https://www.buzzfeednews.com/article/stephaniemlee/coronavirus-antibody-test-santa-clara-los-angeles-stanford

Basic quantities you already know

Exposure or disease classification table:

	True Status of Exposure or Disease	
2-3 Measured Exposure (Dx) Status	Positive	Negative
Positive	a	b
Negative	c	d

- Sensitivity (Se) = a / (a + c)
- Specificity (Sp) = d / (b + d)
- Positive Predictive Value (PPV) = a / (a + b)
- Negative Predictive Value (NPV) = d / (c + d)

Relation between “true” and expected observed data

The number of individuals expected to be observed in a given study is a function of the “true” exposure status and the bias parameters (i.e., sensitivity and specificity):

	True		Expected observed cells	
1-3	E_1	E_0	E_1	E_0
D_1	A	B	$a = A(Se_{D_1}) + B(1 - Sp_{D_1})$	$b = A(1 - Se_{D_1}) + B(Sp_{D_1})$
D_0	C	D	$c = C(Se_{D_0}) + D(1 - Sp_{D_0})$	$d = C(1 - Se_{D_0}) + D(Sp_{D_0})$

Who is in the observed “a” cell? It is a mixture of:

- Truly exposed individuals correctly classified: $A(Se_{D_1})$
- Truly unexposed individuals misclassified as exposed: $B(1 - Sp_{D_1})$

Note that observed data *implicitly* assumes 100% Se and Sp

- Suppose we have 85% Se and 95% Sp, which is non-differential with respect to disease:

	True		Expected observed cells	
1-3	E_1	E_0	E_1	E_0
D_1	200	100	$200(.85) + 100(.05) = 175$	$200(.15) + 100(.95) = 125$
D_0	800	900	$800(.85) + 900(.05) = 725$	$800(.15) + 900(.95) = 975$
	1000	1000	900	1100

The observed a cell has 175 individuals, of which:

- 170 (200×0.85) are truly exposed and correctly classified: $A(Se_{D_1})$
- 5 (100×0.05) are truly unexposed but misclassified: $B(1 - Sp_{D_0})$

$$RR_{true} = \frac{200/1000}{100/900} = 2.0 \qquad RR_{obs} = \frac{175/900}{125/1100} = 1.7$$

Going from observed to “true” data

- We often have “observed” data and want to know what the corrected effect would be.
- Re-arrange the equations above to go from “observed” to “true” cells:
- Recall that the observed a cell is: $a = A(Se_{D_1}) + B(1 - Sp_{D_1})$, and we know that the “true” B cell must be $B = D_{1Tot} - A$
- Now we can substitute: $a = A(Se_{D_1}) + (D_{1Tot} - A)(1 - Sp_{D_1})$ and solve for A

$$
\begin{aligned}
a &= A(Se_{D1}) + D_{1Tot} - A - D_{1Tot}(Sp_{D1}) + A(Sp_{D1}) \\
a - D_{1Tot} + D_{1Tot}(Sp_{D1}) &= A(Se_{D1}) - A + A(Sp_{D1}) \\
a - D_{1Tot}(1 - Sp_{D1}) &= A(Se_{D1} - 1 + Sp_{D1}) \\
\frac{a - D_{1Tot}(1 - Sp_{D1})}{(Se_{D1} - 1 + Sp_{D1})} &= A
\end{aligned}
$$

Going from observed to “true” data

	Observed			Corrected	
1-4	E_1	E_0	Total	E_1	E_0
D_1	a	b	D_{1Tot}	$\frac{a - D_{1Tot}(1 - Sp_{D_1})}{Se_{D_1} - (1 - Sp_{D_1})}$	$D_{1Tot} - A$
D_0	c	d	D_{0Tot}	$\frac{c - D_{0Tot}(1 - Sp_{D_0})}{Se_{D_0} - (1 - Sp_{D_0})}$	$D_{0T} - C$

In our earlier example Se=85% and Sp=95%, we observed $a = 175$, so to get the “true” estimate we have:

$$A = \frac{a - D_{1Tot}(1 - Sp_{D1})}{(Se_{D1} - 1 + Sp_{D1})} = \frac{175 - 300(0.05)}{(0.85 - 1 + 0.95)} = 200$$

What can be done about misclassification?

- Validation study
  - Measurement using a “gold standard” on a random sub-sample.
  - Need to make assumptions about who “complies” with additional measurements and participation.
  - Might still be infeasible to conduct a validation study for other reasons (e.g., data already collected).
  - However, for many kinds of exposures (e.g., Personality tests, social class, etc.) there are no “gold standard” tests.
- One option in these cases is to try and quantify the potential role that misclassification may have played in your study.
- If misclassification is ignored entirely, you are assuming 100% Sensitivity and 100% Specificity.

Implications of validation for bias parameters

Choices for internal validation study:

1. Sample by misclassified exposure and obtain “gold standard”:
    - Directly estimates PPV/NPV, but not Se/SP.
    - Need to consider outcome and confounders.
2. Sample by “true” exposure:
    - Directly estimates Se/Sp but not PPV/NPV.
    - Often not feasible.
3. Sample randomly:
    - Can estimate Se/Sp/PPV/NPV.
    - Need adequate samples for rare exposures or outcomes, potentially stratified by covariates.

Example: non-differential misclassification of exposure

Example of coronary heart disease (D) and "Type A" personality (E):

	True		Observed cells		
1-3	E_1	E_0	E_1	E_0	Total
D_1	A	B	150	107	257
D_0	C	D	1277	1620	2897

$$RR_{obs} = (150/1427)/(107/1727) = 1.7$$

- Suppose we have good reasons for assuming 80% Se and 90% Sp.
- What is the true association?

Calculating “true” values from observed

In our example Se=80% and Sp=90%, we observed $a = 150$, so to get the “true” estimate for A we have:

$$A = \frac{a - D_{1Tot}(1 - Sp_{D1})}{(Se_{D1} - 1 + Sp_{D1})} = \frac{150 - 257(0.10)}{(0.80 - 1 + 0.90)} = 177.6$$

Similar calculations for “true” C:

$$C = \frac{c - D_{0Tot}(1 - Sp_{D0})}{(Se_{D0} - 1 + Sp_{D0})} = \frac{1277 - 2897(0.10)}{(0.80 - 1 + 0.90)} = 1410.4$$

Example: non-differential misclassification of exposure

- True B and D are then calculated by subtraction from marginal totals:

$$B = D_{1Tot} - a = 257 - 177.5 = 79.4$$
$$D = D_{0Tot} - b = 2897 - 1410.4 = 1486.6$$

- Now fill in the “true” table:

	True		**Observed cells**		
1-3	E_1	E_0	E_1	E_0	Total
D_1	177.6	79.4	150	107	257
D_0	1410.4	1486.6	1277	1620	2897

$$RR_{true} = (177.6/1588)/(79.4/1566) = 2.2$$

Simple implementation via episensr[1]

```r
library(episensr)
# observed data
misclassification(
  matrix(c(150, 107,
           1277, 1620),
  dimnames = list(c("D+", "D-"),
  c("E+", "E- ")),
  nrow = 2, byrow = TRUE),
  type = "exposure",

# bias parameters
  bias_parms = c(0.80, 0.80,
                 0.90, 0.90))
```

```
--Observed data--
         Outcome: D+
       Comparing: E+ vs. E-

     E+  E-
D+  150  107
D- 1277 1620

                                          2.5%    97.5%
Observed Relative Risk: 1.696586 1.337392 2.152252
   Observed Odds Ratio: 1.778409 1.373122 2.303319
---
                                                                  2.5%    97.5%
Misclassification Bias Corrected Relative Risk: 2.204640
   Misclassification Bias Corrected Odds Ratio: 2.356302 1.554052 3.572700
```

1. See package by Haine (2021)

Simpler implementation using PPV/NPVs[1]

	Observed		Misclassification adjusted data	
	E_1	E_0	E_1	E_0
D_1	a	b	$a(\text{PPV}_{D1}) + b(1 - \text{NPV}_{D1})$	$D_{1\ \text{Total}} - A$
D_0	c	d	$c(\text{PPV}_{D0}) + d(1 - \text{NPV}_{D0})$	$D_{0\ \text{Total}} - C$
Total	$a + c$	$b + d$	$A + C$	$B + D$

Table 6.11 Example of bias-adjustment for misclassification of BMI category using positive and negative predictive values in a study of the effect of BMI category on early preterm birth.

$\text{PPV}_{D1} = 65\%$, $\text{PPV}_{D0} = 74\%$, $\text{NPV}_{D1} = 100\%$, $\text{NPV}_{D0} = 98\%$	Observed		Adjusted data	
	Underweight	Normal weight	Underweight	Normal weight
Preterm	599	4978	389.0	5188.0
Term	31,175	391,851	29,687.7	393,338.3
OR	1.5		1.0	

1. See Ch.6 in Fox et al. (2022)

Misclassification of confounders

- Even if exposure and disease have 100% Se and 100% Sp, misclassification of confounders can lead to biased estimates.
- If the confounding is strong and the exposure-disease relation is weak or null, confounder misclassification can produce misleading results, even if independent and non-differential.
- Need bias parameters for misclassified confounder data (validation, literature, etc.)

Stratified analysis implementation

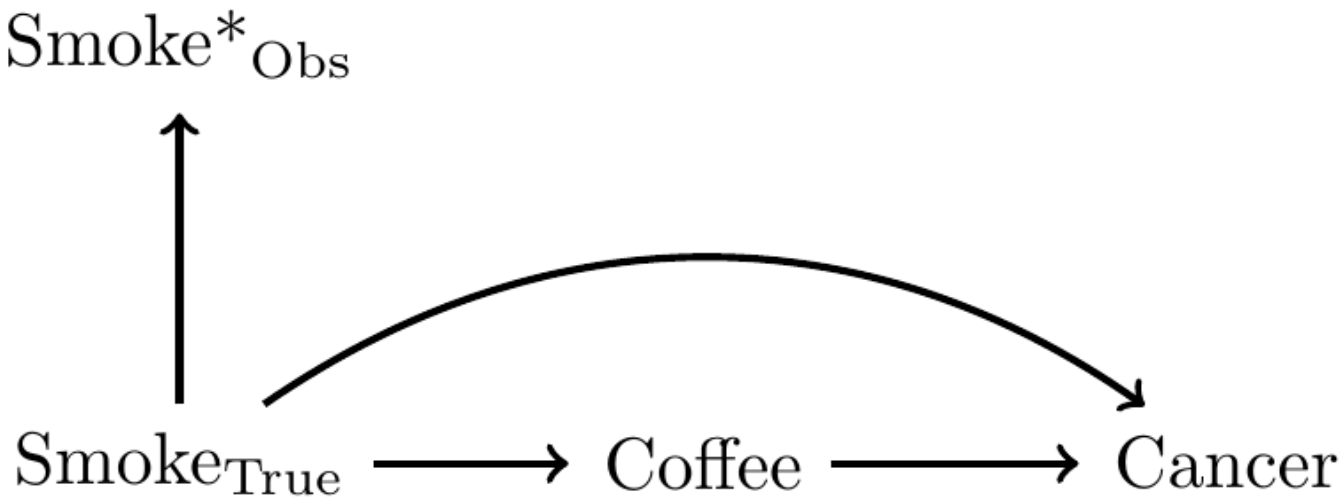
- Estimates for Se and Sp are applied to measurement of confounder rather than exposure.[1]

Observed data						
	Total		C_1		C_0	
	E_1	E_0	E_1	E_0	E_1	E_0
D_1	a	b	a_{C1}	b_{C1}	a_{C0}	b_{C0}
D_0	c	d	c_{C1}	d_{C1}	c_{C0}	d_{C0}
Bias-adjusted data						
			C_1		C_0	
	E_1	E_0	E_1	E_0	E_1	E_0
D_1	A	B	$[a_{C1} - a(1 - \text{SP})]/[(\text{SE} - (1 - \text{SP})]$	$[b_{C1} - b(1 - \text{SP})]/[(\text{SE} - (1 - \text{SP})]$	$a - A_{C1}$	$b - B_{C1}$
D_0	C	D	$[c_{C1} - c(1 - \text{SP})]/[(\text{SE} - (1 - \text{SP})]$	$[d_{C1} - d(1 - \text{SP})]/[(\text{SE} - (1 - \text{SP})]$	$c - C_{C1}$	$d - D_{C1}$
Total			$A_{C1} + C_{C1}$	$B_{C1} + D_{C1}$	$A_{C0} + C_{C0}$	$B_{C0} + D_{C0}$

1. See examples in Ch.6 of Fox et al. (2022).

Example: Effect of coffee consumption on bladder cancer[1]

We might draw a model like this, indicating that true smoking status is unmeasured but may still confound the association between coffee and bladder cancer:



1. Slattery et al. (1988)

Excel resources at https://sites.google.com/site/biasanalysis/

Data (Enter Stratified Coffee-Cancer Data in Blue Cells)

	Total		Smoking +		Smoking -	
	Coffee +	Coffee -	Coffee +	Coffee -	Coffee +	Coffee -
Cancer +	202 (a)	322 (b)	144 (A_{C1})	203 (B_{C1})	58 (A_{C0})	119 (B_{C0})
Cancer -	124 (c)	360 (d)	56 (C_{C1})	98 (D_{C1})	68 (C_{C0})	262 (D_{C0})
Total	326 (m)	682 (n)	200 (M_{C1})	301 (N_{C1})	126 (M_{C0})	381 (N_{C0})

$$OR_{crude} = 1.8 \quad \text{and} \quad OR_{MHadj} = 1.5 \quad 95\%\text{CI} = [1.1, 2.0]$$

Bias Adjusted Crude and Adjusted Measures of Coffee-Cancer Relationship

	Total		Smoking +		Smoking -	
	Coffee +	Coffee -	Coffee +	Coffee -	Coffee +	Coffee -
Cancer +	202	322	180.0 (A_{C1})	253.8 (B_{C1})	22.0 (A_{C0})	68.3 (B_{C0})
Cancer -	124	360	70.0 (C_{C1})	122.5 (D_{C1})	54.0 (C_{C0})	237.5 (D_{C0})
Total	326 (m)	682 (n)	250.0 (M_{C1})	376.3 (N_{C1})	76.0 (M_{C0})	305.8 (N_{C0})

$$OR_{MHadj} = 1.3 \quad 95\%\text{CI} = [0.95, 1.7]$$

Overview

Why Bias Analysis?

Deterministic Bias Analysis

Unmeasured confounding

Misclassification

Selection bias

Multidimensional Bias Analysis

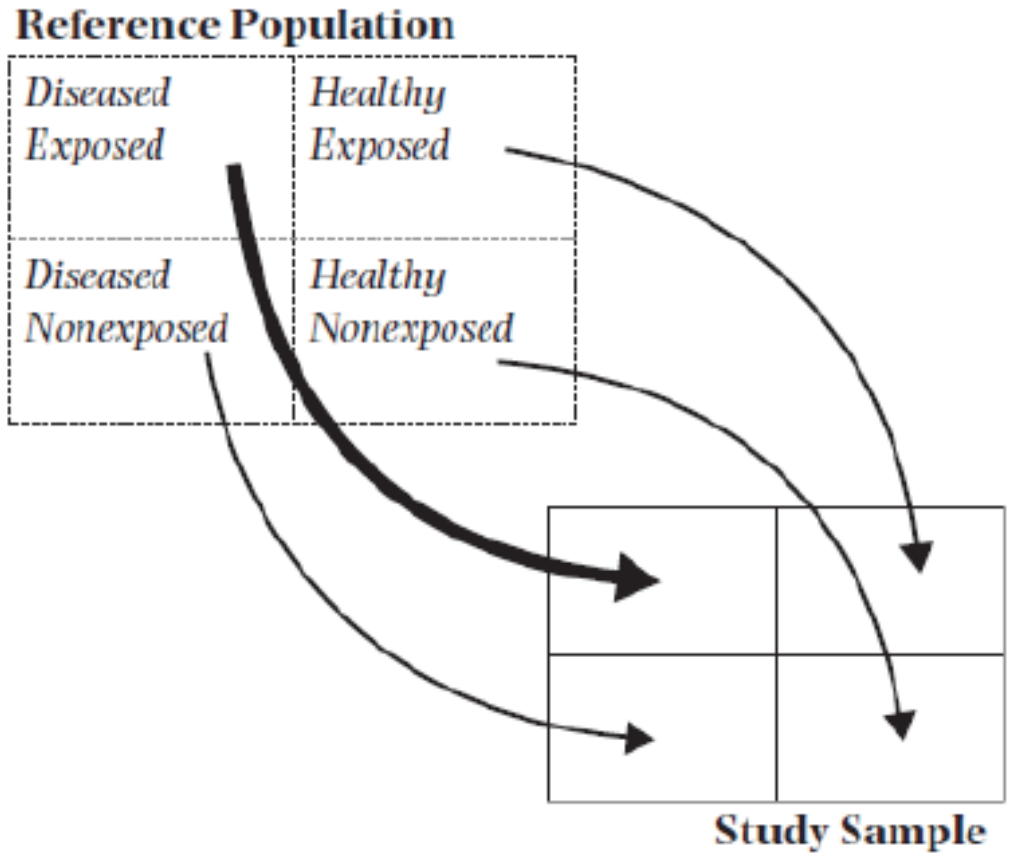
Record-level Implementation

Summary

Bias analysis for selection bias

- Results from conditioning on a common effect of exposure and disease.
  - Differential baseline participation
  - Differential losses to follow-up

Or when the cells of the 2 x 2 table in your study are sampled with different probabilities from the 2 x 2 table in the target population:



Differential baseline participation

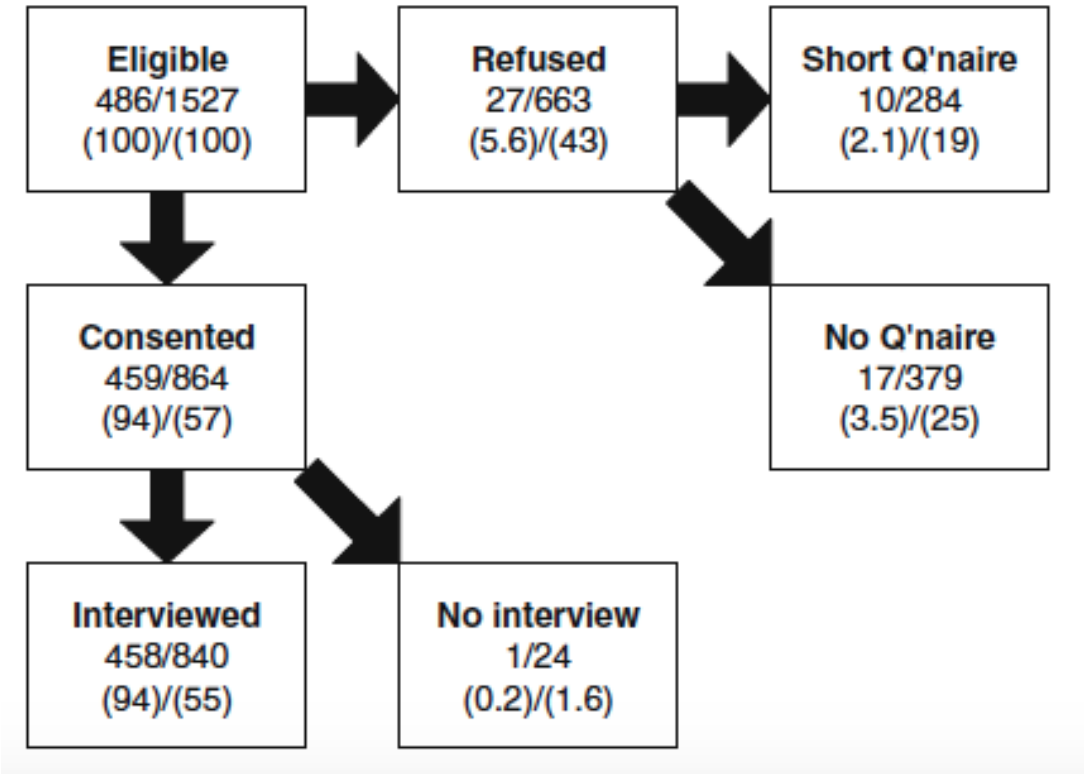
What do you need?

- Selection probabilities
- Ideally, by exposure *and* disease status.

Where can you get it?

- Validation (internal or external).
- Internal probably better.
- Educated guesses (largely based on experience or prior data).

Example: mobile phone use and uveal melanoma[1]



1. Example from Stang et al. (2009) described in Fox et al. (2022)

	Participants		Non-participants+Q		Refusals
2-3	E+	E-	E+	E-	?
Cases	136	107	3	7	17
Controls	297	165	72	212	379

- OR among participants: (136/297)/(107/165) = 0.71
- OR among non-participants w/Q: (3/72)/(7/212) = 1.25
- Who's missing? 3 exposed cases we know from the questionnaire, but what about the other non-participants? Assume the exposure distribution similar to the other non-participants (3/10)*17. Same for controls.
- Now we can get an adjusted OR:

$$OR_{adj} = \frac{136 + 3 + (3/10) * 17}{297 + 72 + (72/284) * 379} / \frac{107 + 7 + (7/10) * 17}{165 + 212 + (212/284) * 379} = 1.63$$

Using selection probabilities

- More generally, if we know or can estimate, or have reasonable guesses about the selection probabilities, we can create an adjusted OR using the selection probabilities:
  - E+ Cases = 136/(136+3+(3/10)*17) = 0.94
  - E- Cases = 107/(107+7+(7/10)*17) = 0.85
  - E+ Controls = 297/(297+72+(72/284)*379) = 0.64
  - E- Controls = 165/(165+212+(212/294)*379) = 0.25

$$OR_{adj} = \hat{OR} \times \frac{S_{caseE-} \times S_{controlE+}}{S_{caseE+} \times S_{controlE-}}$$

$$OR_{adj} = 0.71 \times \frac{0.85 \times 0.64}{0.94 \times 0.25} = 1.63$$

Using inverse-probability of selection weighting

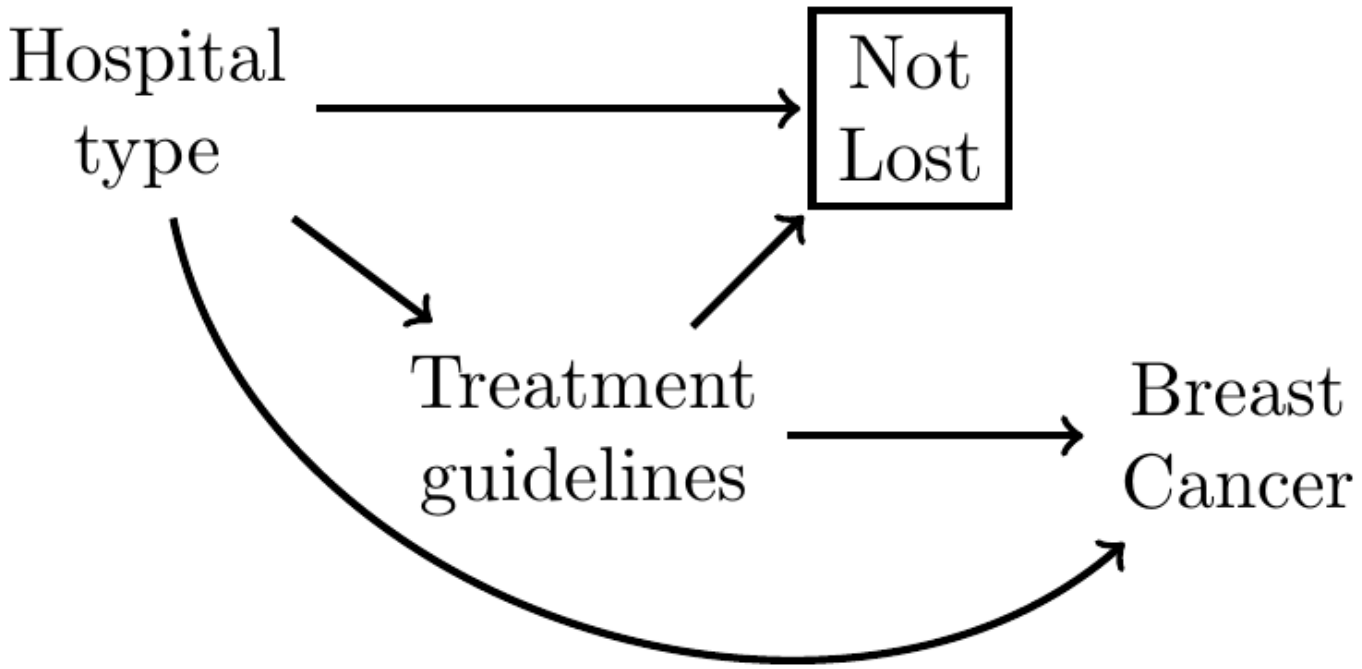
- Could also just reweight each group by the inverse of its probability of selection:
  - Weighted E+ Cases = 136 * (1 / 0.94)
  - Weighted E- Cases = 107 * (1 / 0.85)
  - Weighted E+ Controls = 297 * (1 / 0.64)
  - Weighted E- Controls = 165 * (1 / 0.25)

$$OR_{adj} = \frac{136 * (1/S_{caseE+})/297 * (1/S_{controlE+})}{107 * (1/S_{caseE-})/165 * (1/S_{controlE-})}$$

$$OR_{adj} = \frac{136 * (1/0.94)/297 * (1/0.64)}{107 * (1/0.85)/165 * (1/0.25)} = 1.63$$

Differential loss-to-follow up

- How to account for potential bias among those lost?
- Example of impact of treatment guidelines on breast cancer mortality.
- Overall loss-to-follow of 13%, initial baseline treatment status of those lost is known.



How to estimate the rates among those lost?

	With follow-up		**Lost to follow-up**	
2-3	E+	E-	E+	E-
Deaths	40	65		
Persons	104	286	13	46
Person-years	687	2560		
Crude rate	5.8/100py	2.5/100py		
Crude RD	3.3/100py	0		
Crude RR	2.3	1.0		

- First, let's assume that they would have accrued similar person years as those with follow-up:

	With follow-up		Lost to follow-up	
2-3	E+	E-	E+	E-
Deaths	40	65		
Persons	104	286	13	46
Person-years	687	2560	85.9	411.7
Crude rate	5.8/100py	2.5/100py		
Crude RD	3.3/100py	0		
Crude RR	2.3	1.0		

$$PY_{E+} = (687/104) \times 13 = 85.9py$$
$$PY_{E-} = (2560/286) \times 46 = 411.7py$$

- What should we assume about their mortality risks?
- What if we assume similar to those with follow-up?

- Missing individuals were diagnosed at 2 specific hospitals, so use observed rates among those non-missing in hospitals where those LTF were diagnosed

$$Deaths_{E+} = .049 \times 85.9 = 4.2$$
$$Deaths_{E-} = .045 \times 411.7 = 18.7$$

	At 2 hospitals		Estimated for those lost	
2-3	E+	E-	E+	E-
Deaths	3	5	4.2	18.7
Person-years	60.8	110.2	85.9	411.7
Crude rate	4.9/100py	4.5/100py		
Crude RD	0.4/100py	0		
Crude RR	1.1	1.0		

Bias-corrected estimates (with assumptions)

	With follow-up		Lost to follow-up	
2-3	E+	E-	E+	E-
Deaths	40	65	4.2	18.7
Persons	104	286	13	46
Person-years	687	2560	85.9	411.7
Crude rate	5.8/100py	2.5/100py		
Crude RD	3.3/100py	0		
Crude RR	2.3	1.0		

$$IR_{E+} = (40 + 4.2)/(687py + 85.9py) = 5.7/100py$$
$$IR_{E-} = (65 + 18.7)(2560py + 411.7py) = 2.8/100py$$
$$RD = 2.9/100py \quad \text{and} \quad RR = 2.0$$

Worst case scenario still suggests some impact.

	With follow-up		Lost to follow-up	
2-3	E+	E-	E+	E-
Deaths	40	65	0	46
Persons	104	286	13	46
Person-years	687	2560	85.9	411.7
Crude rate	5.8/100py	2.5/100py		
Crude RD	3.3/100py	0		
Crude RR	2.3	1.0		

$$IR_{E+} = (40 + 0)/(687py + 85.9py) = 5.2/100py$$
$$IR_{E-} = (65 + 46)(2560py + 411.7py) = 3.7/100py$$
$$RD = 1.4/100py \quad \text{and} \quad RR = 1.4$$

IP weighting for selection bias

	Diagnosed at other than two hospitals		Diagnosed at two hospitals	
	< guideline	guideline	< guideline	guideline
Cases in reidentified	37	60	3	5
Reidentified N	96	273	8	13
Reidentified PY	626.2	2449.8	60.8	110.2
Not reidentified N	3	11	10	35
Total N	99	284	18	48
Reidentification proportion	0.97	0.96	0.44	0.27
IPAW	1.03	1.04	2.25	3.69
Crude rate (/100 PY)	5.9	2.4	4.9	4.5
Stratum specific IRD (/100 PY)	3.5		0.4	
Stratum specific IRR	2.4		1.1	
Crude IRD (/100 PY)	3.3			
Crude IRR	2.3			
IPAW IRD (/100 PY)	3.0			
IPAW IRR	2.10			

Original data from Silliman et al., 1989 and Lash et al., 2000 [18, 19].
PY person years, *IPAW* inverse probability of attrition weights, *IRD* incidence rate difference, *IRR* incidence rate ratio.

Break! ☕

07:00

Overview

Why Bias Analysis?

Deterministic Bias Analysis

Unmeasured confounding

Misclassification

Selection bias

Multidimensional Bias Analysis

Record-level Implementation

Summary

Multidimensional bias analysis

- Often, bias parameters are only educated guesses.
- Multidimensional bias analysis uses a range of plausible values for bias parameters.
- Especially useful if there are no known validation data for your parameters of interest.

What to vary?

- Unmeasured confounding
  - Vary strength of Z-Exp and Z-Dis associations.
- Misclassification:
  - Range of Se/Sp values, including differential.
- Selection bias
  - Range of different selection proportions or selection bias factor.

Example

Peri-operative consultation and 30d mortality Confounder-adjusted RR=1.16 (1.07,1.26).[1]

Table 5. Effect of an Unmeasured Confounder on the Estimated Association of Preoperative Medical Consultation With 30-Day Mortality

		Adjusted Association of Consultation With Mortality,[c] OR (95% CI), by Risk Associated With Unmeasured Confounder[d]			
$P_{consult}$, %[a]	$P_{no\ consult}$, %[b]	OR 1.25	OR 1.50	OR 2.00	OR 2.50
0	0.0	1.16 (1.07-1.26)[e]	1.16 (1.07-1.26)[e]	1.16 (1.07-1.26)[e]	1.16 (1.07-1.26)[e]
10	6.7	1.15 (1.06-1.25)	1.14 (1.05-1.24)	1.13 (1.04-1.22)	1.11 (1.02-1.21)
10	5.0	1.15 (1.06-1.24)	1.13 (1.04-1.23)	1.11 (1.02-1.20)	1.08 (1.01-1.17)
20	13.3	1.14 (1.05-1.24)	1.12 (1.04-1.22)	1.10 (1.01-1.19)	1.07 (0.99-1.16)
20	10.0	1.13 (1.05-1.23)	1.11 (1.02-1.20)	1.06 (0.98-1.15)	1.03 (0.95-1.11)
30	20.0	1.13 (1.05-1.23)	1.11 (1.02-1.21)	1.07 (0.99-1.16)	1.04 (0.96-1.13)
30	15.0	1.12 (1.03-1.22)	1.08 (1.00-1.18)	1.03 (0.95-1.11)	0.98 (0.90-1.06)
40	26.7	1.12 (1.04-1.22)	1.10 (1.01-1.19)	1.05 (0.97-1.14)	1.02 (0.94-1.10)
40	20.0	1.11 (1.02-1.20)	1.06 (0.98-1.15)	0.99 (0.92-1.08)	0.94 (0.87-1.02)

> An unmeasured confounder could render the association between consultation and 30-day mortality statistically nonsignificant but only if it at least doubled the odds of mortality and was present in 20% of patients who underwent consultation as compared with 10% of those who did not.

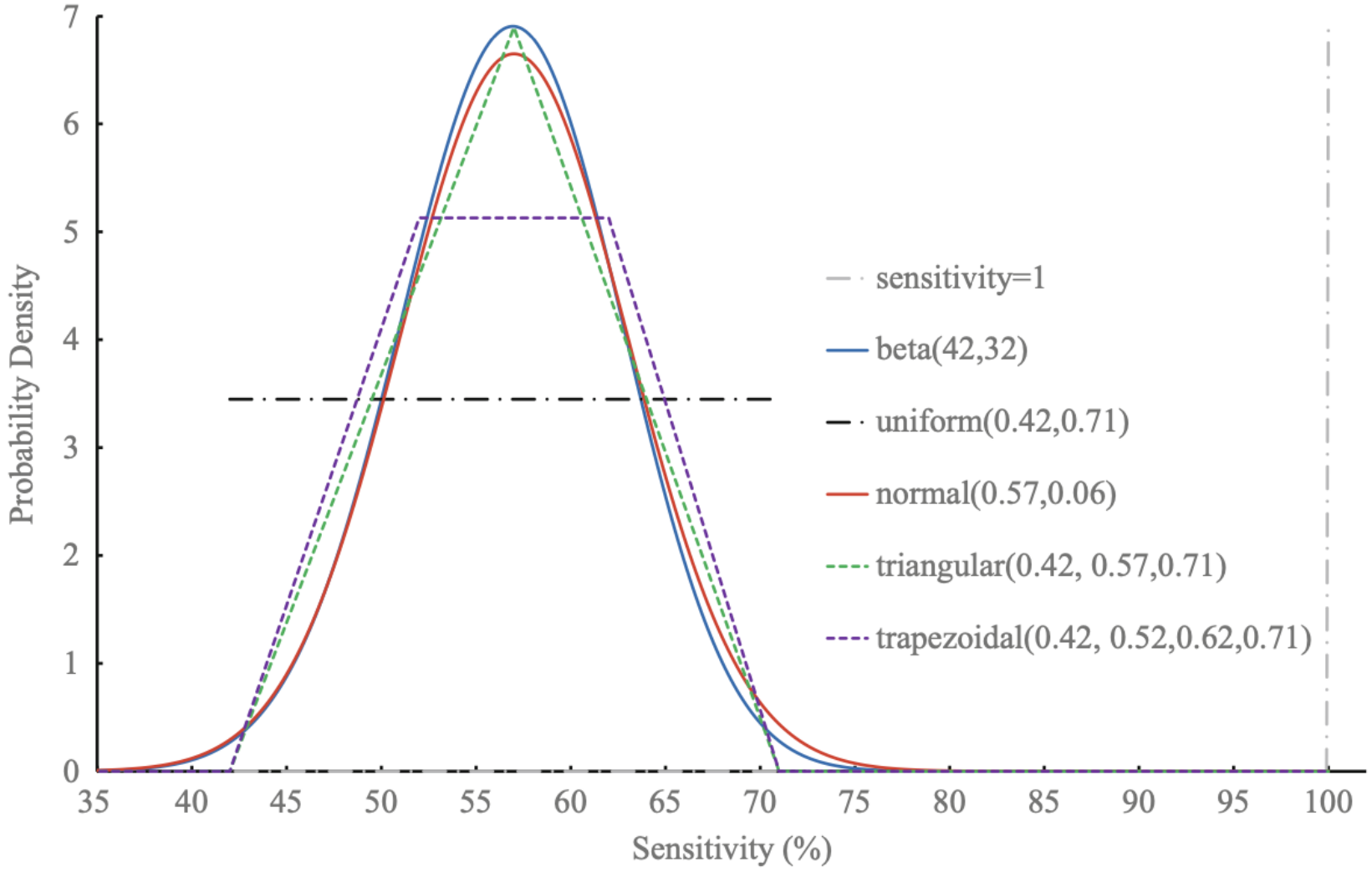
1. Wijeysundera et al. (2010)

Probabilistic Bias Analysis

- The major limitation of the previous methods for analyzing study bias is that they treat the bias parameters (sensitivity, specificity, confounder-disease association, prevalence of unmeasured confounders, etc.) as known quantities that are perfectly measured.
- Thus, the analyses above are referred to as deterministic, and they only account for systematic error (i.e., they do not account for measurement error in the estimates of the prevalence of the unmeasured confounder or the confounder-disease association.
- An alternative would be to assign a probability distribution to bias parameters.

Example distributions for sensitivity

Five distributions all centered near 0.57 and spread around 0.42 and 0.71.

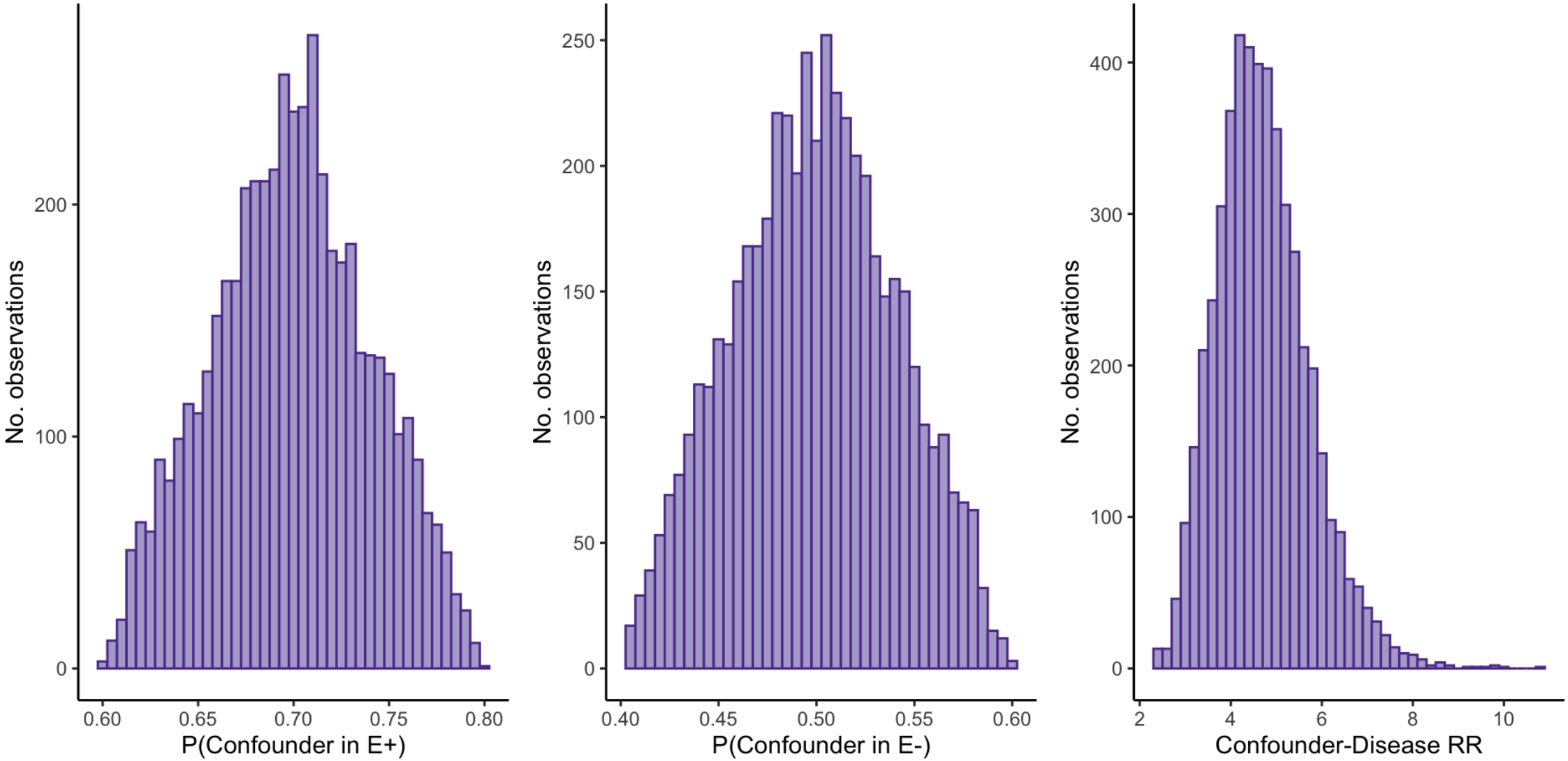


Example: Unmeasured confounding of resins and cancer

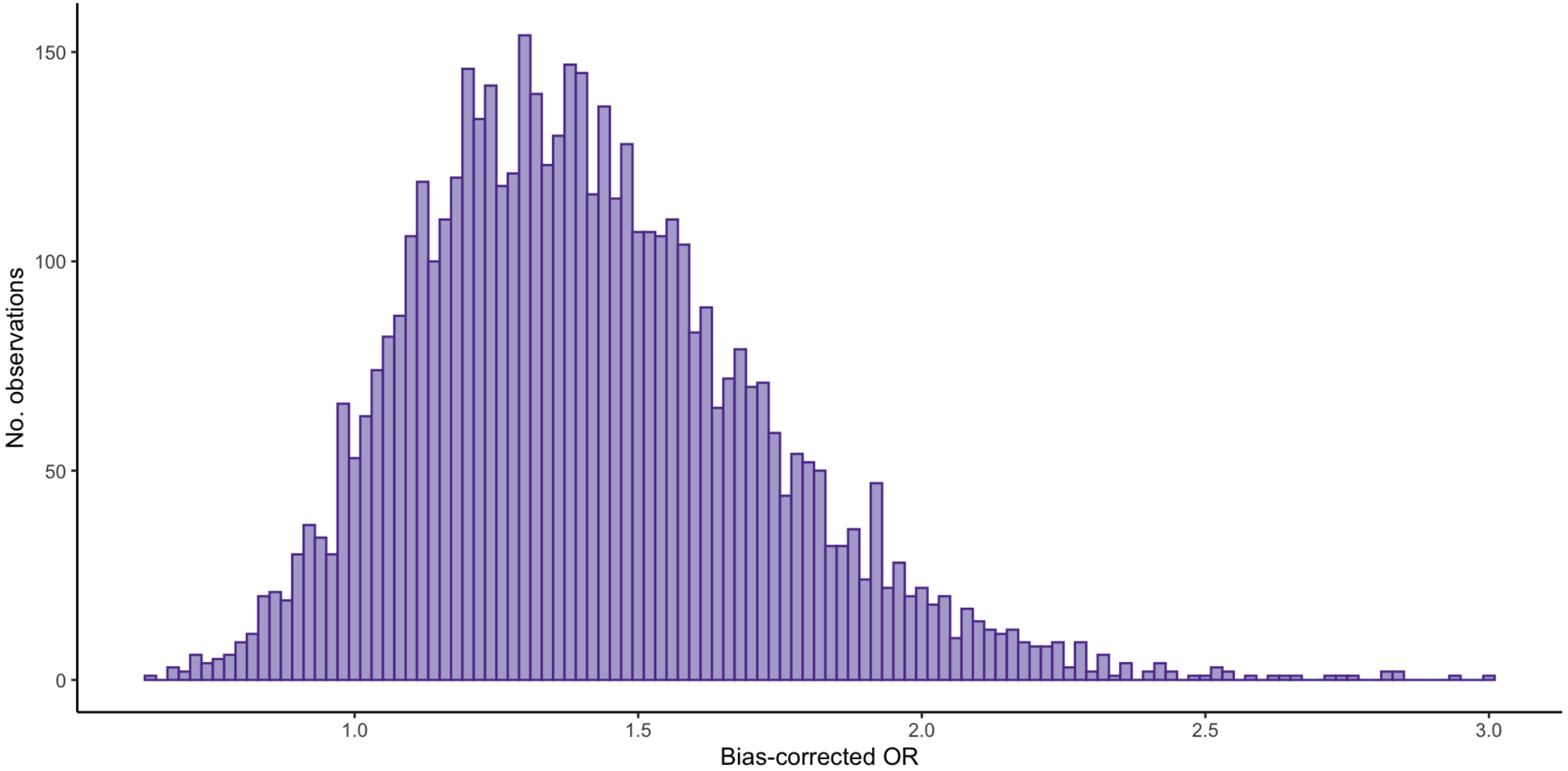
- Setup using the `episensr` package

```r
# set the seed for reproducible results
set.seed(39569)

# define the observed 2x2 data
pc <- probsens.conf(matrix(c(45, 94, 257, 945),
  dimnames = list(c("D+", "D-"), c("D+", "D-")), nrow = 2, byrow = TRUE),

# number of replications
  reps = 5000,

# bias parameters and distributions

# prevalence of U among exposed
prev.exp = list("triangular", c(.6, .8, .7)),

# prevalence of U among unexposed
prev.nexp = list("triangular", c(.4, .6, .5)),

# association of U with outcome
risk = list("log-normal", c(1.522, 0.216)),

# correlation between exposure prevalences
corr.p = 0.01)
```

Probability distributions for each parameter



Generates a distribution of corrected estimates



Bias-corrected estimates

	Median	2.5th pctile	97.5th pctile
OR:Crude	1.76	1.20	2.58
OR:Corrected-systematic error	1.38	1.17	1.61
OR:Corrected-systematic and random error	1.38	0.90	2.10

- Using probabilistic analysis gives a similar adjusted OR (1.44) relative to our deterministic bias analysis (1.39), but provides empirical confidence limits.
- Introduction of additional random error via the simulation by choosing a standard normal deviate (z_i) for each iteration (i), multiply by SE

$$estimate_i^{total} = estimate_i^{adj} - z_i \times SE_i^{adj}$$

$$OR_i^{total} = e^{ln(OR_i^{adj}) - z_i \times SE_i^{adj}}$$

Multiple bias analysis

- Extension of simple bias analysis in which we assign bias parameters, either deterministically or probabilistically, but now we examine the impact of more than one bias at a time.
- Difficult to ascertain quantitatively how multiple biases may work together.
- Order matters, and corrections should be made in the reverse of the order in which they occurred as the data were generated.
- Generally (but not a rule):
  1. Misclassification
  2. Selection bias
  3. Unmeasured confounding

Multiple bias analysis example[1]

Associations between prepregnancy obesity and cleft lip with or without cleft palate, adjusting for different combinations of biases, National Birth Defects Prevention Study, 1997–2011

Model	Conventional analysis		Nonprobabilistic bias analysis*	Probabilistic bias analysis†		Probabilistic bias analysis plus random error†	
	OR	95% CI	OR	Median OR	95% SI	Median OR	95% RESI
Unadjusted	1.09	0.97, 1.21					
Confounding only‡	1.10	0.98, 1.23					
Selection bias and confounding‡			0.98	0.98	0.82, 1.17	0.98	0.80, 1.21
Exposure misclassification and confounding‡,§							
Nondifferential ($Se_{ca} = Se_{co}$, $Sp_{ca} = Sp_{co}$)			1.12	1.11	0.84, 1.45	1.11	0.83, 1.48
Differential A ($Se_{ca} = Se_{co}$+0.05, $Sp_{ca} = Sp_{co}$+0.03)			1.09	1.09	0.84, 1.41	1.09	0.82, 1.44
Differential B ($Se_{ca} = Se_{co}$-0.05, $Sp_{ca} = Sp_{co}$-0.03)			1.02	1.01	0.76, 1.34	1.01	0.74, 1.36
All biases combined‡,‖							
Nondifferential ($Se_{ca} = Se_{co}$, $Sp_{ca} = Sp_{co}$)			1.03	1.02	0.76, 1.37	1.02	0.75, 1.39
Differential A ($Se_{ca} = Se_{co}$+0.05, $Sp_{ca} = Sp_{co}$+0.03)			1.01	1.01	0.76, 1.33	1.00	0.74, 1.36
Differential B ($Se_{ca} = Se_{co}$-0.05, $Sp_{ca} = Sp_{co}$-0.03)			0.94	0.93	0.69, 1.26	0.93	0.67, 1.28

Se_{ca}, sensitivity for cases; Se_{co}, sensitivity for controls; Sp_{ca}, specificity for cases; Sp_{co}, specificity for controls.
* Fixed values of bias parameters chosen for the analysis.
† Triangular distributions of bias parameters sampled over 5000 iterations.
‡ Adjusted for confounding by maternal race/ethnicity.
§ Adjusted for exposure misclassification and missing exposure data.
‖ Adjusted for selection bias, exposure misclassification, missing exposure, and confounding.

1. Johnson et al. (2018)

Overview

Why Bias Analysis?

Deterministic Bias Analysis

Unmeasured confounding

Misclassification

Selection bias

Probabilistic Bias Analysis

Record-level Implementation

Summary

Record level implementation

- Applications to summaries have limitations (stratified analyses)
- Often need QBA on estimates adjusted for multiple covariates

Benefits

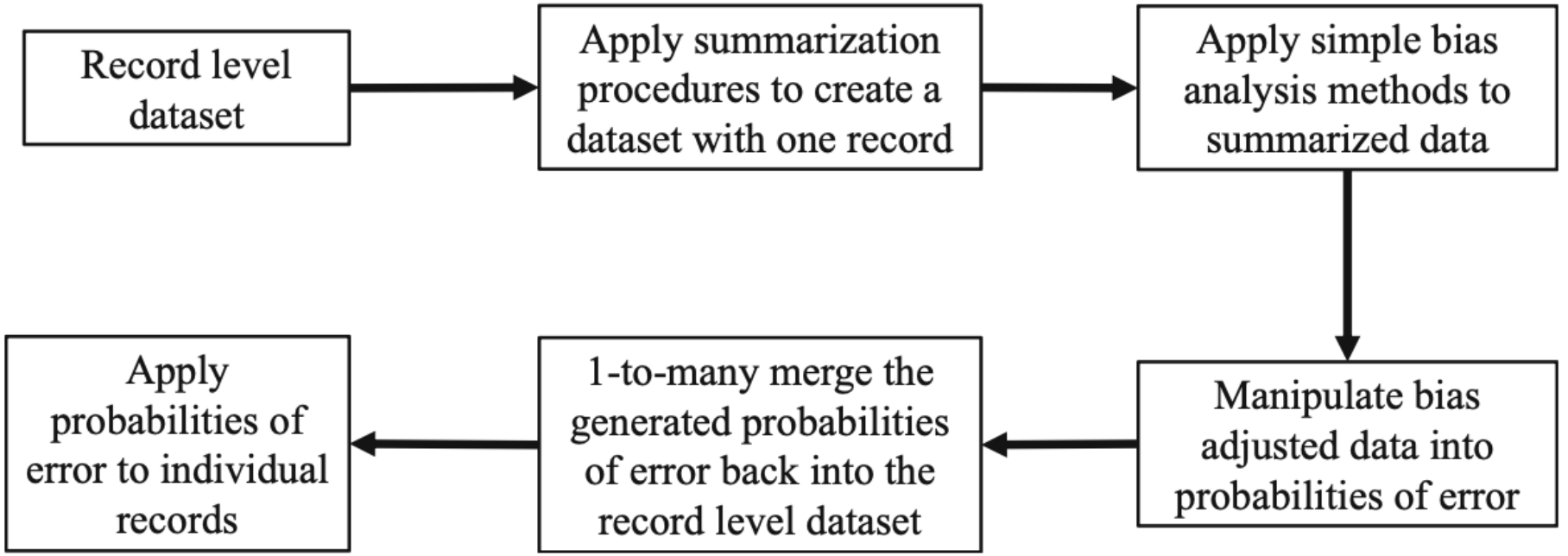
- More complex models (regression)
- More ‘realistic’

Drawbacks

- More complex programming
- Added computing time
- 100K simulations on 10K obs = 100m records

Basic workflow for record-level analysis[1]

- Basic idea is to take *one* realization of bias parameters, apply it to record-level data, then save estimates and repeat.



1. See Fox et al. (2022) Chapter 9.

Steps for record-level implementation

- Step 1: Identify the Source of Bias
- Step 2: Select the Bias Parameters
- Step 3: Assign Probability Distributions to Each Bias Parameter
- Step 4: Use Simple Bias Analysis Methods to Incorporate Uncertainty in the Bias Parameters and Random Error
  - Step 4a: Randomly Sample from the Bias Parameter Distributions
  - Step 4b: Use Simple Bias Analysis Methods and Incorporate Uncertainty and Conventional Random Error
  - Step 4c: Sample the Bias-Adjusted Effect Estimate
- Step 5: Save the Bias-Adjusted Estimate and Repeat Steps 4a–c

Example for Unmeasured Confounding

- Step 3: Assign Probability Distributions to Each Bias Parameter
- Choose triangular distributions for P_1, P_0, and RR_{cd}, which leads to the following realizations for a single set of parameters:

$$RR_{cd} = 0.567$$

$$P_1 = 75.1\%$$

$$P_0 = 7.2\%$$

Collapse record-level data to contingency table, i.e.,

From

Figure 9.9 First ten observations for a record-level dataset for use in a record-level bias analysis for an uncontrolled confounder of the association between male circumcision (e) and HIV (d). Original data from Tyndall et al., 1996 [4].

id	e	d
1	1	0
2	1	0
3	0	0
4	1	0
5	1	0
6	1	0
7	1	0
8	0	1
9	1	1
10	0	0

To:

id	a	b	c	d
1	105	85	527	93

Perform simple QBA:[1]

Table 9.12 Data on the association between male circumcision (E) and HIV (D) stratified by an unmeasured confounder, religious category when $RR_{CD} = 0.567$, $p_1 = 75.1\%$, and $p_0 = 7.2\%$ (C; C_1 – Muslim, C_0 – any other religion).

	Total		C_1		C_0	
	E_1	E_0	E_1	E_0	E_1	E_0
D_1	105	85	66.3	3.6	38.7	81.4
D_0	527	93	408.4	9.2	118.6	83.8
Total	632	178	474.6	12.8	157.4	165.2

Crude data from Tyndall et al., 1996 [4].

iter	rr.cd	p1	p0	M1	M0	N1	N0	A1	B1	C1	D1	A0	B0	C0	D0
1	0.567	0.751	0.072	474.6	157.4	12.8	165.2	66.3	3.6	408.4	9.2	38.7	81.4	118.6	83.8
2	0.587	0.783	0.039	494.9	137.1	6.9	171.1	71.3	2.0	423.6	4.9	33.7	83.0	103.4	88.1
3	0.623	0.779	0.058	492.0	140.0	10.3	167.7	72.1	3.1	420.0	7.2	32.9	81.9	107.0	85.8
4	0.533	0.875	0.078	553.3	78.7	13.8	164.2	82.9	3.7	470.4	10.2	22.1	81.3	56.6	82.8
5	0.627	0.783	0.045	494.6	137.4	7.9	170.1	72.8	2.4	421.8	5.5	32.2	82.6	105.2	87.5
6	0.696	0.807	0.036	510.0	122.0	6.4	171.6	78.1	2.1	431.8	4.2	26.9	82.9	95.2	88.8
7	0.571	0.867	0.061	547.6	84.4	10.8	167.2	82.7	3.0	464.9	7.8	22.3	82.0	62.1	85.2
8	0.732	0.865	0.075	546.5	85.5	13.3	164.7	86.5	4.8	460.0	8.6	18.5	80.2	67.0	84.4
9	0.650	0.761	0.075	481.2	150.8	13.3	164.7	70.8	4.3	410.4	9.1	34.2	80.7	116.6	83.9
10	0.777	0.833	0.077	526.2	105.8	13.7	164.3	83.4	5.2	442.8	8.5	21.6	79.8	84.2	84.5

1. Note that the values in the table are in the first observation row.

Generate predicted probabilities for the unmeasured confounder:

Parameter	Estimate
pr(C + |E + D+)	63.1%
pr(C + |E-D+)	4.2%
pr(C + |E + D-)	77.5%
pr(C + |E-D-)	9.9%

iter	rr.cd	p1	p0	prc.e1d1	prc.e0d1	prc.e1d0	prc.e0d0
1	0.567	0.751	0.072	0.631	0.042	0.775	0.099
2	0.555	0.816	0.048	0.711	0.027	0.837	0.067
3	0.569	0.878	0.043	0.804	0.025	0.893	0.060
4	0.673	0.800	0.071	0.729	0.049	0.814	0.091
5	0.652	0.838	0.047	0.771	0.031	0.851	0.061

- Merge back into the *record-level* data
- Generate the confounder via a Bernoulli trial based on the predicted probability for each cell. Something like `c=rbinom(n,1,p)`.
- If the trial returns a 1, set $c = 1$, if not, then $c = 0$, i.e., does not hav the confounder.

id	e	d	prc.e1d1	prc.e0d1	prc.e1d0	prc.e0d0	p	c
1	1	0	0.631	0.042	0.775	0.099	0.775	1
2	1	0	0.631	0.042	0.775	0.099	0.775	1
3	0	0	0.631	0.042	0.775	0.099	0.099	0
4	1	0	0.631	0.042	0.775	0.099	0.775	0
5	1	0	0.631	0.042	0.775	0.099	0.775	1
6	1	0	0.631	0.042	0.775	0.099	0.775	0
7	1	0	0.631	0.042	0.775	0.099	0.775	0
8	0	1	0.631	0.042	0.775	0.099	0.042	0
9	1	1	0.631	0.042	0.775	0.099	0.631	1
10	0	0	0.631	0.042	0.775	0.099	0.099	0

- Estimate the parameter for each iteration, save, and summarize:

iter	rr.cd	p1	p0	rr_reg	se_reg	rr_tot
1	0.567	0.751	0.072	0.531	0.142	0.580
2	0.583	0.826	0.055	0.551	0.163	0.585
3	0.582	0.792	0.066	0.533	0.152	0.482
4	0.711	0.884	0.059	0.335	0.225	0.282
5	0.506	0.753	0.077	0.555	0.142	0.566
6	0.615	0.830	0.041	0.473	0.165	0.535
7	0.701	0.766	0.056	0.448	0.163	0.466
8	0.520	0.863	0.052	0.557	0.177	0.561
9	0.543	0.804	0.077	0.528	0.161	0.472
10	0.588	0.777	0.089	0.506	0.153	0.502

Table 9.15 Results of a record-level probabilistic bias analysis of the relationship between male circumcision and HIV adjusting for an unmeasured confounder, religious category.

	Record-level analysis		Summary-level analysis[a]	
Analysis	Median	95% Interval	Median	95% Interval
Conventional result, assuming no bias	0.348	(0.276, 0.439)	0.348	(0.276, 0.439)
Systematic error only[a]	0.472	(0.418, 0.550)	0.472	(0.419, 0.549)
Random and systematic error	0.471	(0.300, 0.722)	0.474	(0.293, 0.726)

Summary

- Bias analysis is a method for trying to quantify *systematic* sources of bias not captured in sampling error.
- Generally, don't rely on intuitions, especially the idea that "non-differential" bias will lead to bias toward the null.
- Consider implementing validation studies to help ground your choice of bias parameters.
- Be humble and honest about uncertainty.
- Likely that simple, plausible, deterministic QBA is better than nothing.

Is it worth it?

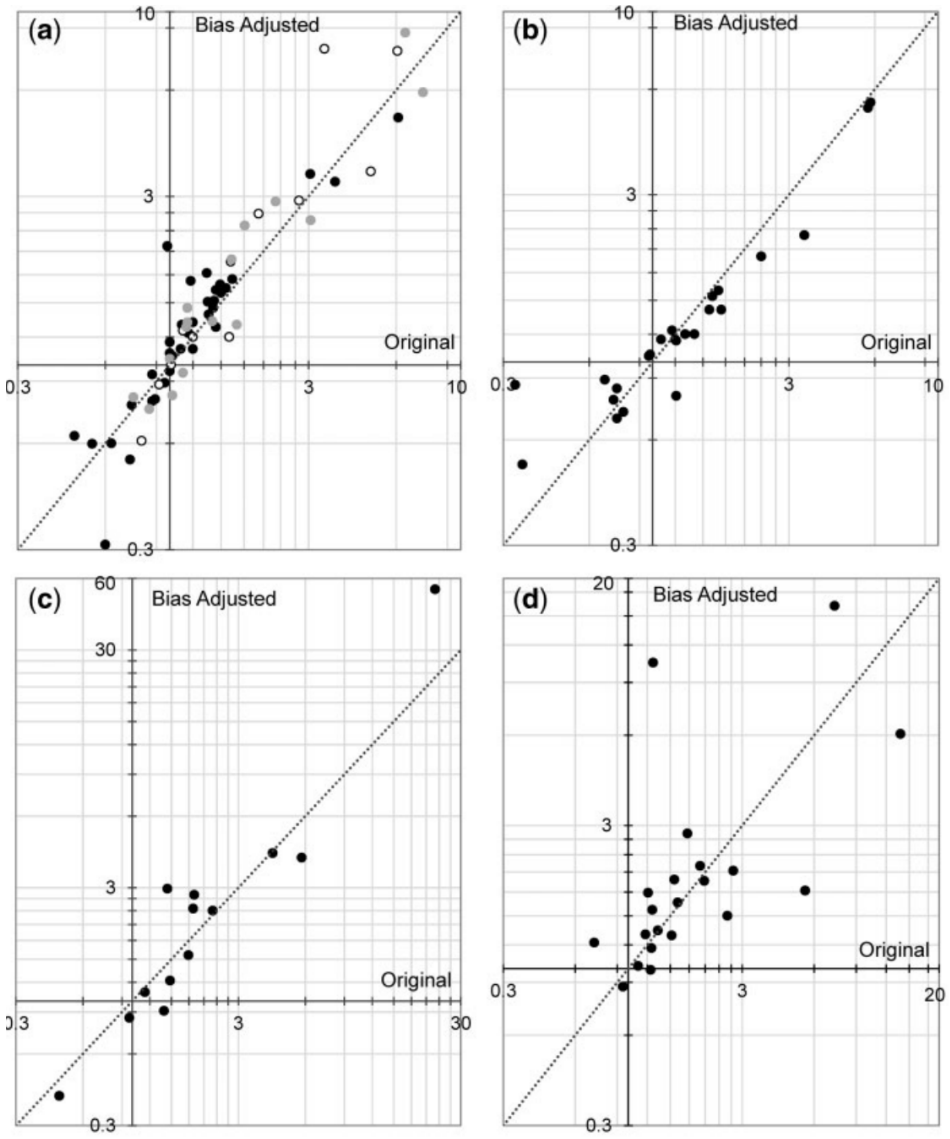
Results of systematic review of QBAs.[1] Graph shows original and revised estimates for:

a. misclassification

b. unmeasured confounding

c. selection bias

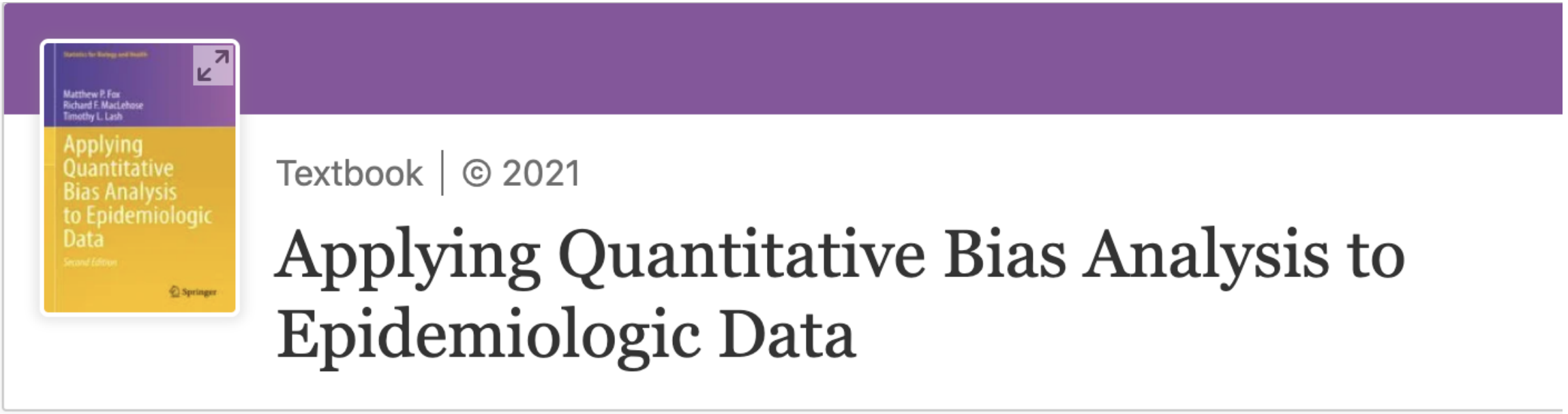
d. multiple biases



1. Petersen et al. (2021)

Best practices

Try to get through all of this:



References

Arah OA, Chiba Y, Greenland S. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. Ann Epidemiol. 2008 Aug;18(8):637–46.

Bross I. Misclassification in 2 x 2 tables. Biometrics. 1954;10(4):478–86.

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: Recent evidence and a discussion of some questions. J Natl Cancer Inst. 1959 Jan;22(1):173–203.

Epidemiology Monitor. Gary taubes faces epidemiology [published 1996] [Internet]. [cited 2011]. Available from: http://www.digitalsmarttools.com/eEpiMon/Interview_Taubes.htm

Fox MP, MacLehose RF, Lash TL. Applying quantitative bias analysis to epidemiologic data. Springer; 2022.

Greenland S, Salvan A, Wegman DH, Hallock MF, Smith TJ. A case-control study of cancer mortality at a transformer-assembly facility. International archives of occupational and environmental health. 1994;66:49–54.

Haine D. The episensr package: Basic sensitivity analysis of epidemiological results [Internet]. 2021. Available from: https://dhaine.github.io/episensr/

Johnson CY, Howards PP, Strickland MJ, Waller DK, Flanders WD, National Birth Defects Prevention Study. Multiple bias analysis using logistic regression: An example from the national birth defects prevention study. Ann Epidemiol. 2018 Aug;28(8):510–4.

McCullough LE, Maliniak ML, Amin AB, Baker JM, Baliashvili D, Barberio J, et al. Epidemiology beyond its limits. Sci Adv. 2022 Jun;8(23):eabn3328.

Parekh N, Chappell RJ, Millen AE, Albert DM, Mares JA. Association between vitamin d and age-related macular degeneration in the third national health and nutrition examination survey, 1988 through 1994. Arch Ophthalmol. 2007 May;125(5):661–9.

Petersen JM, Ranker LR, Barnard-Mayers R, MacLehose RF, Fox MP. A systematic review of quantitative bias analysis applied to epidemiological research. Int J Epidemiol. 2021 Nov;50(5):1708–30.

Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society: Series B (Methodological). 1983;45(2):212–8.