Simple sensitivity analysis for selection bias using bounds

Louisa H. Smith

Harvard T.H. Chan School of Public Health

December 19, 2020

CMStatistics 2020

What is selection bias?

Sensitivity analysis for selection bias using bounds

Everything is *selection bias*

- Model selection
  - Problems with statistical inference
- Confounding
  - In certain fields… “selection into treatment”
- Non-generalizability/transportability
  - Magnitude of effect in sample not the same as in target population
- **Collider stratification**
  - Bias for causal effects even within sample and under the null

[Smith, 2020]

Selection bias in this talk

A: exposure of interest (binary for simplicity)
Y: binary outcome of interest
S: indicator of selection into study ($S = 1$ if selected, $S = 0$ if eligible but no data)

We can estimate an observed risk ratio

$$\text{RR}_{AY}^{obs} = \frac{\Pr(Y = 1 \mid A = 1, S = 1)}{\Pr(Y = 1 \mid A = 0, S = 1)}$$

which may not equal the true causal risk ratio RR_{AY}^{true}.

- We will assume that if we could estimate $\frac{\Pr(Y=1|A=1)}{\Pr(Y=1|A=0)}$, we would be estimating RR_{AY}^{true}.

Example from Hernán et al., 2004

Consider a randomized trial of anti-retroviral therapy (A) among people living with HIV, with a goal of preventing the development of AIDS (Y)

- $\frac{\Pr(Y=1|A=1)}{\Pr(Y=1|A=0)}$ is the risk ratio among people randomized to the intervention arm vs. standard of care
- If some people drop out of the study, we can only estimate $\frac{\Pr(Y=1|A=1,S=1)}{\Pr(Y=1|A=0,S=1)}$

I'll use boxes around nodes on graphs to indicate conditioning on those nodes.

Why might bias arise?

Those eligible for the study are not a random sample of all people living with HIV... is that a problem?

- Perhaps, if we're trying to estimate how effective treatment would be in another context.

- This is a problem of generalizeability / transportability (external validity)

But *not* when it comes to estimating valid causal effects.

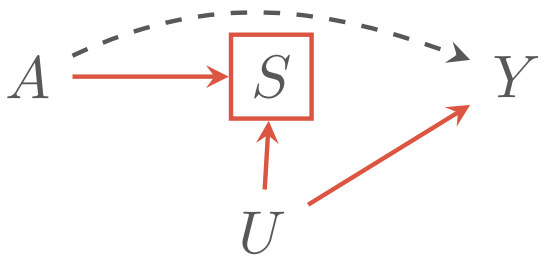
- With complete follow-up, we can estimate the effect of the drug in the target population from which the participants came.

- With loss to follow-up, we can't even estimate that (no internal validity).
  - Not even if we only want to infer things about the people for whom $S = 1$

Why not?

The participants who were lost to follow-up are *not* a random sample of all participants

- Perhaps the most severely immunocompromised (U) people have trouble coming to study visits
- They are also at higher risk of developing AIDS
- Perhaps people experiencing side effects of treatment no longer want to participate



Common structure

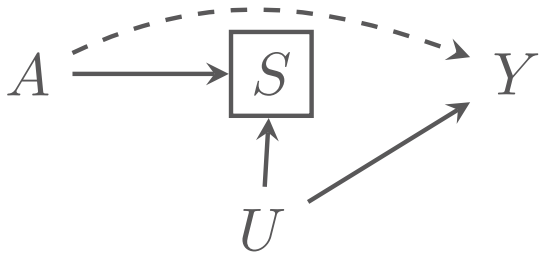
Does Zika virus infection (A) increase the risk of microcephaly (Y)?

- We only assess microcephaly among live births ($S = 1$).
- Elective terminations are not included ($S = 0$).

Is the selected group different?

We might assume that

- People who have more exposure to the virus are more likely to choose to end their pregnancies (worried about risks)
- People with less access to health care are less likely to have access to abortion services
- There are factors that affect risk of microcephaly that are correlated with access to health care



- The pregnancies most likely to *not* be terminated are those at risk of microcephaly for other reasons
  - It looks like exposure to Zika virus is associated with increased risk of microcephaly

A note about confounding

- There are also confounders of the A - Y relationship, of course, since this study is observational
- Some of those might be the same factors causing selection bias
  - If we properly adjust for them to control confounding, we also control selection bias
- If there are additional factors leading to selection bias that aren't confounders, we may not plan to measure or adjust for them
- We'll assume confounders are measured (in which case we're estimating RR_{AY}^{obs} within strata) or controlled by study design
  - Everything conditional on $C = c$

What is selection bias?

Sensitivity analysis for selection bias using bounds

When you run into selection bias

I think there's selection bias in this study

When you run into selection bias

I think there's selection bias in this study

→ Fix it!

(Recruit different participants, track down lost to follow-up)

When you run into selection bias

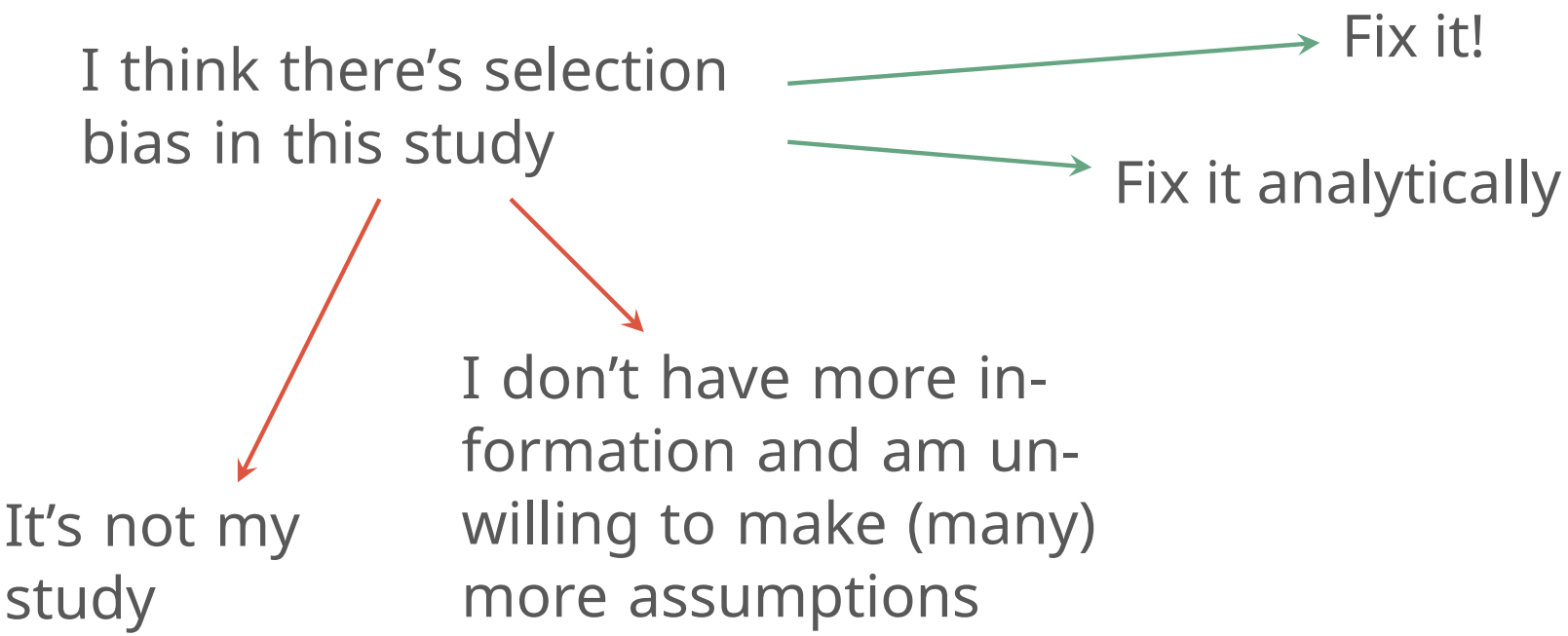
I think there's selection bias in this study

→ Fix it!

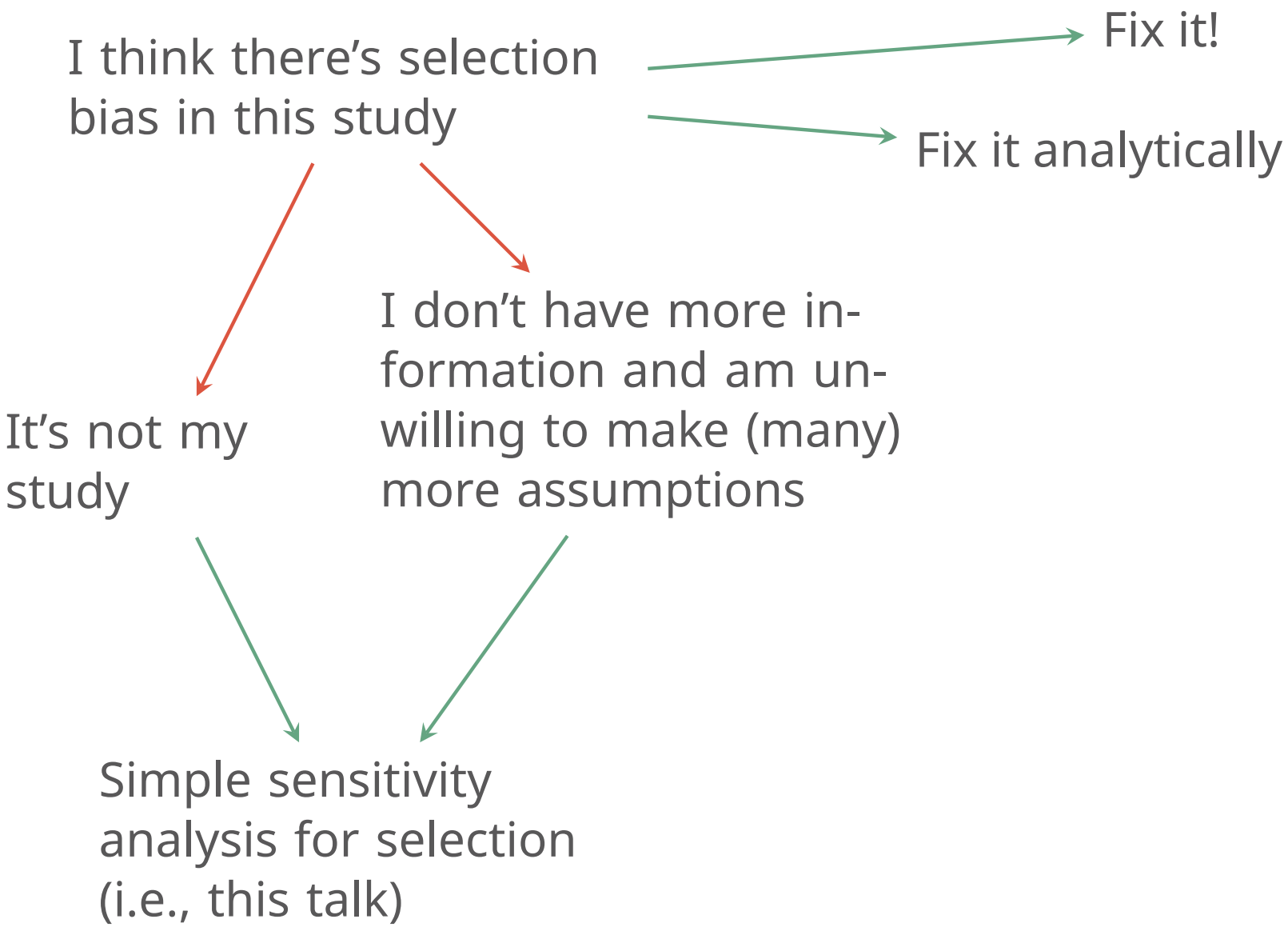
→ Fix it analytically

(Measure and model)

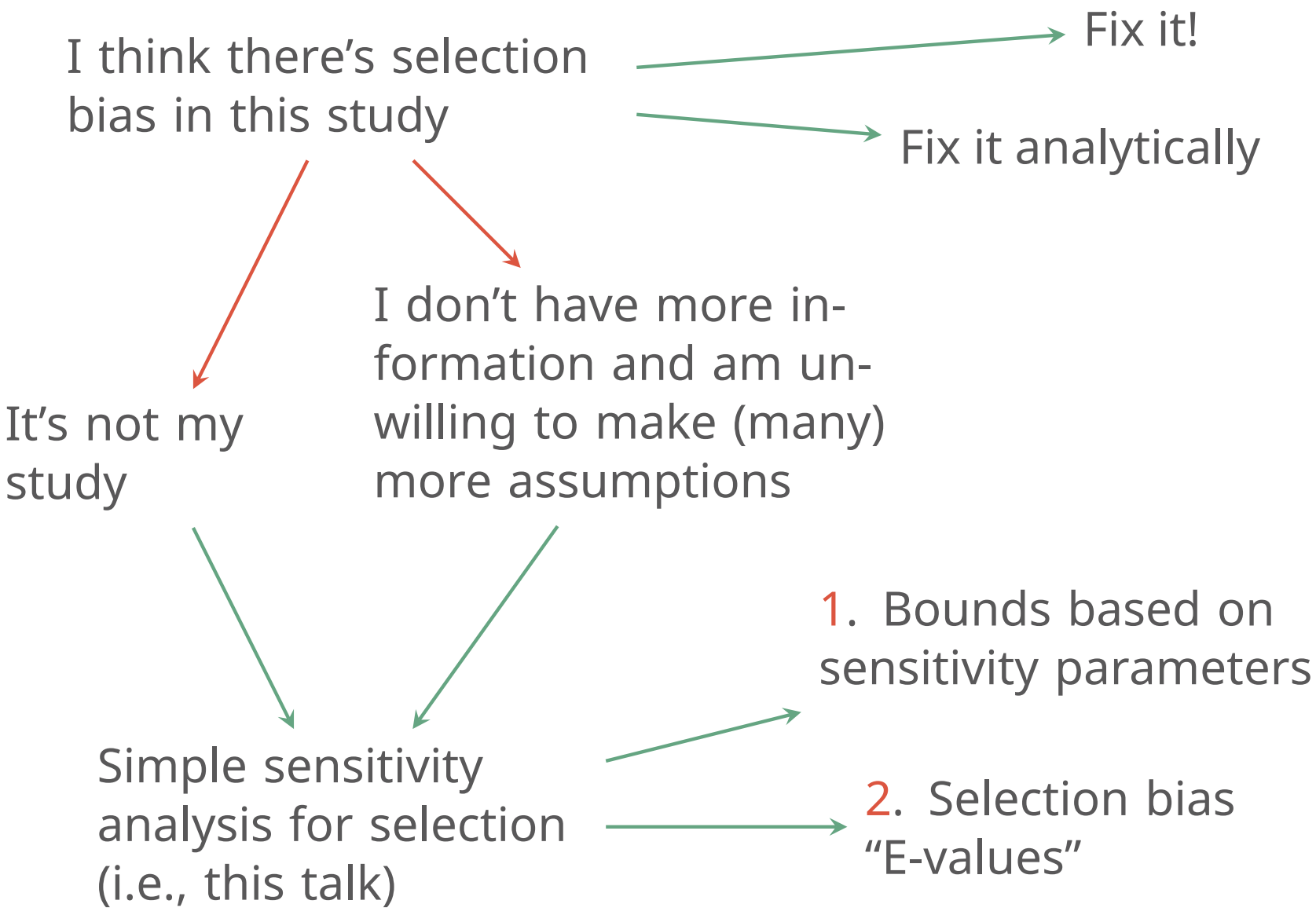
When you run into selection bias



When you run into selection bias



When you run into selection bias



Framework for sensitivity analysis

Define the relative bias:

$$\text{bias} = \text{RR}_{AY}^{obs} \big/ \text{RR}_{AY}^{true}$$

strength of selection $X \implies$

$$\text{bias} \leq f(X) \implies$$

$$\text{RR}_{AY}^{true} \geq \text{RR}_{AY}^{obs} \big/ f(X)$$

If A is protective (RR < 1), invert everything

1. Selection bias bounds

Propose values for X...

... and use $f(X)$ to "correct" the observed risk ratio (conservatively)

$$\text{RR}_{AY}^{true} \geq \text{RR}_{AY}^{obs} \Big/ f(X)$$

2. Selection bias ``E-values''

What if the true causal effect were null?

$$\text{bias} = \text{RR}_{AY}^{obs} / 1$$
$$\text{bias} \leq f(X)$$
$$\text{RR}_{AY}^{obs} \leq f(X)$$

Then the minimum strength of selection, in terms of X, that could result in that much bias:

$$X \geq f^{-1}(\text{RR}_{AY}^{obs})$$

We can also consider non-null true effects.

Problem solved for unmeasured confounding

Define sensitivity parameters in terms of unmeasured confounder(s) U

- Ding and VanderWeele 2016; VanderWeele and Ding 2017

RESEARCH AND REPORTING METHODS Annals of Internal Medicine

Sensitivity Analysis in Observational Research: Introducing the E-Value

Tyler J. VanderWeele, PhD, and Peng Ding, PhD

Sensitivity analysis is useful in assessing how robust an association is to potential unmeasured or uncontrolled confounding. This article introduces a new measure called the "E-value," which is related to the evidence for causality in observational studies that are potentially subject to confounding. The E-value is defined as the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment–outcome association, conditional on the measured covariates. A large E-value implies that considerable unmeasured confounding would be needed to explain away an effect estimate. A small E-value implies little unmeasured confounding would be needed to explain away an effect estimate. The authors propose that in all observational studies intended to produce evidence for causality, the E-value be reported or some other sensitivity analysis be used. They suggest calculating the E-value for both the observed association estimate (after adjustments for measured confounders) and the limit of the confidence interval closest to the null. If this were to become standard practice, the ability of the scientific community to assess evidence from observational studies would improve considerably, and ultimately, science would be strengthened.

Ann Intern Med. 2017;167:268-274. doi:10.7326/M16-2607 **Annals.org**
For author affiliations, see end of text.
This article was published at Annals.org on 11 July 2017.

Solving the problem for selection bias

What do X (the sensitivity parameters) and $f(X)$ (the bound) need to look like to use either bounds or E-values as a sensitivity analysis?

- A little more complicated than unmeasured confounding
- It depends on the target population, the structure of the selection bias, other assumptions you're willing to make

1. Bound for inference in the whole population

If the structure of selection bias is such that $Y \perp\!\!\!\perp S \mid A, U$:

$$\text{bias} \leq \left(\frac{\text{RR}_{UY|(A=1)} \times \text{RR}_{SU|(A=1)}}{\text{RR}_{UY|(A=1)} + \text{RR}_{SU|(A=1)} - 1} \right) \times \left(\frac{\text{RR}_{UY|(A=0)} \times \text{RR}_{SU|(A=0)}}{\text{RR}_{UY|(A=0)} + \text{RR}_{SU|(A=0)} - 1} \right)$$

- $\text{RR}_{UY|(A=a)} = \dfrac{\max u \, \Pr(Y = 1 \mid A = a, U = u)}{\min u' \, \Pr(Y = 1 \mid A = a, U = u')}$
- $\text{RR}_{SU|(A=a)} = \max u \dfrac{\Pr(U = u \mid A = a, S = s)}{\Pr(U = u \mid A = a', S = s')}$

1. Bound for inference in the whole population

The sensitivity parameters answer the questions:

- $\text{RR}_{UY|(A=a)}$: To what extent is the outcome risk increased by the unmeasured factor, within a single level of the exposure?

- $\text{RR}_{SU|(A=a)}$: To what extent is some value of the unmeasured factor more prevalent among the selected compared to the non-selected group?

Zika virus example: bound

- Suppose that lack of access to medical care was associated with 2-fold higher risk of microcephaly among both the Zika-exposed and unexposed (conditional on measured factors)
  - $\text{RR}_{UY|(A=1)} = \text{RR}_{UY|(A=0)} = 2$
- Suppose that lack of access to medical care for pregnant women was up to 1.7 times more likely for women *without* an induced abortion among the Zika-exposed
  - $\text{RR}_{SU|(A=1)} = 1.7$
- Suppose that access to medical care was up to 1.5 times more likely for women with an induced abortion among the unexposed
  - $\text{RR}_{SU|(A=0)} = 1.5$

Zika virus example: bound

Plugging in these plausible values, we have

$$\left(\frac{\mathsf{RR}_{UY|(A=1)} \times \mathsf{RR}_{SU|(A=1)}}{\mathsf{RR}_{UY|(A=1)} + \mathsf{RR}_{SU|(A=1)} - 1}\right) \times \left(\frac{\mathsf{RR}_{UY|(A=0)} \times \mathsf{RR}_{SU|(A=0)}}{\mathsf{RR}_{UY|(A=0)} + \mathsf{RR}_{SU|(A=0)} - 1}\right) =$$

$$\left(\frac{2 \times 1.7}{2 + 1.7 - 1}\right) \times \left(\frac{2 \times 1.5}{2 + 1.5 - 1}\right) = 1.51$$

Zika virus example: bound

- From de Araújo et al. (2018) we have $\widehat{\text{RR}}_{AY}^{obs} = 73.1$ for the Zika-microcephaly relationship with a lower confidence limit of 13.0.

- If our hypothesized values are true, we know that the maximum selection bias would be a factor of 1.51.
  - We can "correct" the point estimate and lower confidence limit: 73.1 / 1.51 = 48.1 and 13.0 / 1.51 = 8.6.
  - Under our assumptions, the true causal effect estimate must be **at least** of that magnitude.

2. E-value for selection bias

The observed risk ratio could be fully explained by selection bias if, for $a = 0, 1$:

$$\text{RR}_{UY|(A=a)} = \text{RR}_{SU|(A=a)} \geq \sqrt{\text{RR}_{AY}^{obs}} + \sqrt{\text{RR}_{AY}^{obs} - \sqrt{\text{RR}_{AY}^{obs}}}$$

- This is a way to summarize the minimal "strength" of selection bias that could explain away a result

Zika virus example: E-value

$$\sqrt{73.1} + \sqrt{73.1 - \sqrt{73.1}} = 16.6$$

If

$$\mathsf{RR}_{UY|(A=0)} = \mathsf{RR}_{UY|(A=1)} = \mathsf{RR}_{SU|(A=0)} = \mathsf{RR}_{SU|(A=1)} \geq 16.6$$

it is possible that there is no causal Zika-microcephaly relationship and the observed risk ratio was entirely due to selection bias

- Worst-case scenario

Extensions

- Different bound if only wish to make inference about the selected group
- Assumptions about the directionality of the bias
- Some results on the risk difference scale
- Bound for selection bias and unmeasured confounding (and misclassification)

[Smith and VanderWeele, 2019; Smith, Mathur, et al., 2020]

Software

Implemented in the R package EValue

```
library(EValue)
multi_bound(selection(), RRUsYA1 = 2, RRSUsA1 = 1.7,
            RRUsYA0 = 2, RRSUsA0 = 1.5)
```

```
## [1] 1.511111
```

```
multi_evalue(selection(), OR(73.1, rare = TRUE), lo = 13)
```

```
##                          point     lower upper
## RR                   73.10000 13.000000    NA
## Multi-bias E-values 16.58415  6.670587    NA
```

Naturally extends to additional biases

[Mathur et al., 2018; Smith, Mathur, et al., 2020]

selection-bias.louisahsmith.com

Simple sensitivity analysis for selection bias | Introduction | Compute bound | E-values for se

Outcome type

Risk ratio

Target population

- (•) Entire population
- ( ) Selected population

Necessary assumptions

- [x] No unmeasured confounding $(Y_a \amalg A \mid C)$
- [x] Selection is only related to outcome via unmeasured factor(s) U $(Y \amalg S \mid A, U, C)$

Additional assumptions

- [ ] Unmeasured factor a defining characteristic of selection
- [ ] Selection always associated with increased risk of outcome in both exposure groups
- [ ] Selection always associated with decreased risk of outcome in both exposure groups

Estimated/hypothesized values for parameters

$\text{RR}_{UY\mid(A=0)}$	$\text{RR}_{UY\mid(A=1)}$	$\text{RR}_{SU\mid(A=0)}$	$\text{RR}_{SU\mid(A=1)}$

Please enter values for the parameters above

Acknowledgements and contact

Thanks to my coauthors:

- Tyler VanderWeele
- Maya Mathur

louisa_h_smith@g.harvard.edu

@louisahsmith

References

- Smith LH. Selection Mechanisms and Their Consequences: Understanding and Addressing Selection Bias. Current Epidemiology Reports 2020.
- Hernán MA, Hernández-Díaz S, and Robins JM. A structural approach to selection bias. Epidemiology 2004;15:615–25.
- Ding P and VanderWeele TJ. Sensitivity Analysis Without Assumptions. Epidemiology 2016;27:368–77.
- VanderWeele TJ and Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. Annals of Internal Medicine 2017;167:268–75.
- de Araújo TVB, Ximenes RAdA, Miranda-Filho DdB, et al. Association between microcephaly, {Zika} virus infection, and other risk factors in {Brazil}: final report of a case-control study. Lancet Infectious Diseases 2018;18:328–36.
- Smith LH and VanderWeele TJ. Bounding Bias Due to Selection. Epidemiology 2019;30:509–16.

References

- Smith LH, Mathur MB, and VanderWeele TJ. Multiple-bias sensitivity analysis using bounds. 2020.
- Mathur MB, Ding P, Riddell CA, et al. Website and R Package for Computing E-values. Epidemiology 2018;29:e45–e47.