UPPSALA UNIVERSITET

Bounding the selection bias

Stina Zetterström and Ingeborg Waernbaum

Department of Statistics, Uppsala University

2023-03-30

UPPSALA UNIVERSITET

Introduction

- The target in many studies is to estimate a causal effect.
- Observational studies are an option when randomized trials are not applicable.
- Two common types of biases in observational studies are:
  - confounding bias
  - selection bias
- Selection bias can arise from missing data, or when the study population is constructed.
  - Study population is often formed by inclusion or exclusion criteria.

UPPSALA UNIVERSITET

"Data and study population"

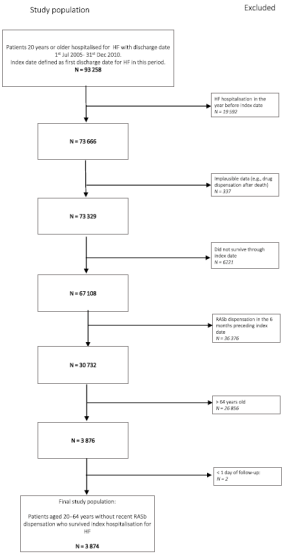
Using this registry, we identified 2 201 352 women who had a first delivery during 1973-2015. To improve internal comparability, only singleton deliveries were included in the analyses, given the higher prevalence of adverse pregnancy outcomes and different underlying causes in multiple gestation pregnancies. We excluded 401 ($\leq 0.1\%$) women with a previous diagnosis of ischemic heart disease and 5 685 (0.3%) women with missing information for pregnancy duration or infant birth weight, leaving 2 195 266 women (99.7% of the original cohort) for inclusion in the study.

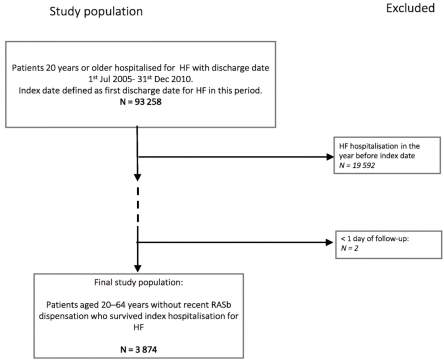
BMJ February 2023

UPPSALA UNIVERSITET

- Model and selection bias
- SV bound
- AF bound
- Comparisons
- Conclusions

"Data and study population"





UPPSALA UNIVERSITET

Our contributions

In this work, we:

- Investigate a bound proposed by Smith and VanderWeele (SV) bound under multiple selections. Smith and VanderWeele (2019)
- Derive results on variation independence and sharpness of the SV bounds.
- Suggest an assumption-free bound.
- Present an R package for calculating these two bounds for selection bias.

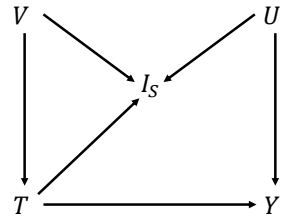
UPPSALA UNIVERSITET

Outline

Model and notation

Variables in the model:

- Binary treatment variable, T.
- Binary potential outcomes, $Y(t),\ t = 0, 1$.
- K binary selection variables, $S_1, ...S_k, ...S_K$.
- Indicator variable constructed from the selection variables, $I_S = \prod S_k$.
- Vector of unmeasured variables, U.
- Vector of observed pre-treatment covariates, X.



Example structure.

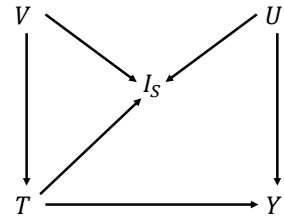
Model and notation

Variables in the model:

- Binary treatment variable, T.
- Binary potential outcomes, $Y(t),\ t = 0, 1$.
- K binary selection variables, $S_1, ...S_k, ...S_K$.
- Indicator variable constructed from the selection variables, $I_S = \prod S_k$.
- Vector of unmeasured variables, U.
- Vector of observed pre-treatment covariates, X.

Assumptions:

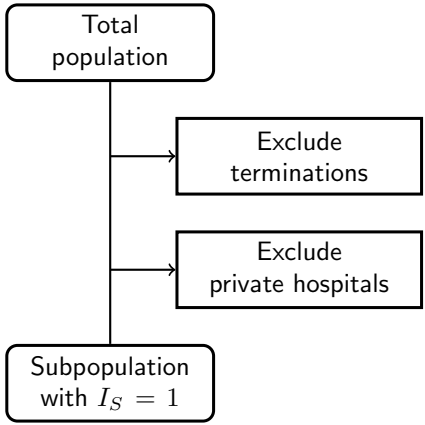
- Consistency, $Y = T \cdot Y(1) + (1 - T) \cdot Y(0)$.
- Conditional exchangeability, $Y(t) \perp\!\!\!\perp T|X,\ t = 0, 1$.
- All analysis is done within stratum $X = x$.
- Ignore sampling variability.

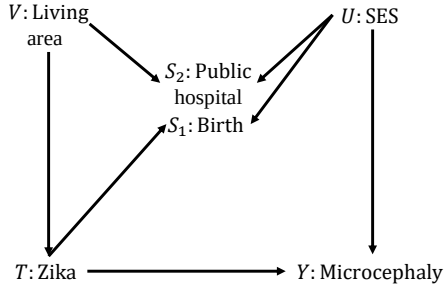


Example structure.

Numerical example

For the purpose of illustration of the bounds we construct a simulated dataset `zika_learner` Smith and VanderWeele (2019), de Araújo et al. (2018)





Causal estimands

- Causal relative risk and causal risk difference in total population:

$$\beta_R = \frac{P(Y(1)=1)}{P(Y(0)=1)}, \quad \beta_D = P(Y(1)=1) - P(Y(0)=1)$$

.

- Causal relative risk and causal risk difference in the selected population:

$$\beta_{R_S} = \frac{P(Y(1)=1|I_S=1)}{P(Y(0)=1)|I_S=1)}, \quad \beta_{D_S} = P(Y(1)=1|I_S=1) - P(Y(0)=1|I_S=1)$$

.

In the `zika_learner`: $\beta_R = 90.7$, $\beta_D = 0.33$, $\beta_{R_S} = 88.1$ and $\beta_{D_S} = 0.36$

Observed estimands

Under selection, $I_S = 1$ we define the observed estimands β_R^{obs} and β_D^{obs} defined as

$$\beta_R^{obs} = \frac{P(Y = 1|T = 1, I_S = 1)}{P(Y = 1|T = 0, I_S = 1)},$$

$$\beta_D^{obs} = P(Y = 1|T = 1, I_S = 1) - P(Y = 1|T = 0, I_S = 1).$$

Ignoring sampling variability = the observed means are treated as an approximation of the corresponding asymptotic mean:

$$\frac{1}{n} \sum_{i:T=t, I_S=1} Y_i \xrightarrow{p} P(Y = 1|T = t, I_S = 1).$$

In the `zika_learner`: $\beta_R^{obs} = 74.5$ and $\beta_D^{obs} = 0.28$

Selection bias

The selection bias is defined as a ratio for the relative risks and as a difference for the risk differences.

$$Bias(\beta_R) = \frac{\beta_R^{obs}}{\beta_R} = \frac{\frac{P(Y=1|T=1,I_S=1)}{P(Y=1|T=0,I_S=1)}}{\frac{P(Y(1)=1)}{P(Y(0)=1)}}$$

$$\begin{aligned} Bias(\beta_D) &= \beta_D^{obs} - \beta_D \\ &= P(Y=1|T=1,I_S=1) - P(Y=1|T=0,I_S=1) \\ &\quad - [P(Y(1)=1) - P(Y(0)=1)] \end{aligned}$$

In the `zika_learner`: $Bias(\beta_R) = 74.5/90.7$ and $Bias(\beta_D) = 0.28 - 0.33$.

Selection bias

Similar for the subpopulation:

$$Bias(\beta_{R_S}) = \frac{\beta_R^{obs}}{\beta_{R_S}} = \frac{\frac{P(Y=1|T=1,I_S=1)}{P(Y=1|T=0,I_S=1)}}{\frac{P(Y(1)=1|I_S=1)}{P(Y(0)=1|I_S=1)}}$$

$$\begin{aligned} Bias(\beta_{D_S}) &= \beta_D^{obs} - \beta_{D_S} \\ &= P(Y=1|T=1,I_S=1) - P(Y=1|T=0,I_S=1) \\ &\quad - [P(Y(1)=1|I_S=1) - P(Y(0)=1|I_S=1)] \end{aligned}$$

In the `zika_learner`: $Bias(\beta_{R_S}) = 74.5/88.1$ and $Bias(\beta_{D_S}) = 0.28 - 0.36$.

Causal estimands unknown $\Rightarrow$ bias unknown $\Rightarrow$ desirable to bound the bias.

UPPSALA
UNIVERSITET

Bounds

Two main approaches when constructing bounds.

1. Make additional assumptions about the causal structure and strengths of the dependencies.
2. Base the bounds solely on the observed data.

Two different bounds are discussed in this work, one from each approach.

UPPSALA UNIVERSITET

Bounds

Two main approaches when constructing bounds.

1. Make additional assumptions about the causal structure and strengths of the dependencies.
2. Base the bounds solely on the observed data.

Two different bounds are discussed in this work, one from each approach.

The bias is bounded from above. If the causal estimand is underestimated, this technicality is solved by recoding the treatment.

UPPSALA UNIVERSITET

SV bound

The first bound is proposed by Smith and VanderWeele, the SV bound.

$\mathcal{B}() \geq Bias()$ using values of sensitivity parameters (based on subject matter/previous knowledge).

For the causal relative risk

$$\mathcal{B}(\beta_R) \geq \beta_R^{obs}/\beta_R \tag{1}$$

and we conclude that the causal relative risk is at least $\beta_R^{obs}/\mathcal{B}(\beta_R)$.

For the causal risk difference the bound $\mathcal{B}(\beta_D)$

$$\mathcal{B}(\beta_D) \geq \beta_D^{obs} - \beta_D \tag{2}$$

and the causal risk difference is at least $\beta_D^{obs} - \mathcal{B}(\beta_D)$.

Conditional independence assumptions

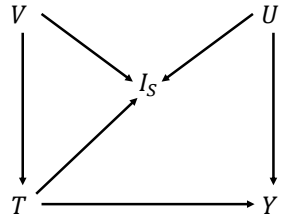
A requirement for the SV bound is an unmeasured variable, U, such that a conditional independence assumption is fulfilled:

Assumption 1

(Total population estimands β_R and β_D) For some unmeasured variable(s) U: $Y \perp\!\!\!\perp I_S | (T = t, U = u)$, *for* $t = 0, 1$.

Assumption 2

(Subpopulation estimands β_{R_S} and β_{D_S}) For some unmeasured variable(s) U: $Y(t) \perp\!\!\!\perp T | (I_S = 1, U = u)$, *for* $t = 0, 1$.



Sensitivity parameters

The sensitivity parameters describe the strength between U and Y, T, and I_S.

Total population estimands: β_R, β_D	Subpopulation estimands: β_{R_S}, β_{D_S}
$RR_{UY\|T=1}=\frac{\max_u P(Y=1\|T=1,U=u)}{\min_u P(Y=1\|T=1,U=u)}$	$RR_{UY\|S=1}=\max_t \frac{\max_u P(Y=1\|T=t,U=u,I_S=1)}{\min_u P(Y=1\|T=t,U=u,I_S=1)}$
$RR_{UY\|T=0}=\frac{\max_u P(Y=1\|T=0,U=u)}{\min_u P(Y=1\|T=0,U=u)}$	
$RR_{SU\|T=1}=\max_u \frac{P(U=u\|T=1,I_S=1)}{P(U=u\|T=1,I_S=0)}$	$RR_{TU\|S=1}=\max_u \frac{P(U=u\|T=1,I_S=1)}{P(U=u\|T=0,I_S=1)}$
$RR_{SU\|T=0}=\max_u \frac{P(U=u\|T=0,I_S=0)}{P(U=u\|T=0,I_S=1)}$	

The sensitivity parameters are unknown and must be guessed.

Which values should be considered?

UPPSALA UNIVERSITET

Feasible regions and variation independence

Only values within the feasible regions should be considered for the SV bounds to be valid. The sensitivity parameters can be restricted by

1. their functional form,
2. each other,
3. the observed data.

Sjölander (2020) investigates similar properties for bounds for confounding bias.

UPPSALA UNIVERSITET

Feasible regions and variation independence

Theorem 1

$\{RR_{UY|T=1}, RR_{UY|T=0}, RR_{SU|T=1}, RR_{SU|T=0}\}$ are restricted by their definitions to values equal to or above 1. Furthermore, for the distribution $P(Y,T,U,I_S)$, there exists a U such that $\{RR_{UY|T=1}, RR_{UY|T=0}, RR_{SU|T=1}, RR_{SU|T=0}\}$ are not restricted by each other or by the observed data distribution, $P(Y,T,I_S)$.

UPPSALA UNIVERSITET

Feasible regions and variation independence

Theorem 1

$\{RR_{UY|T=1}, RR_{UY|T=0}, RR_{SU|T=1}, RR_{SU|T=0}\}$ are restricted by their definitions to values equal to or above 1. Furthermore, for the distribution $P(Y,T,U,I_S)$, there exists a U such that $\{RR_{UY|T=1}, RR_{UY|T=0}, RR_{SU|T=1}, RR_{SU|T=0}\}$ are not restricted by each other or by the observed data distribution, $P(Y,T,I_S)$.

- The analyst can consider all values of the total population sensitivity parameters above or equal to 1 as possible.
- Similar theorem for the sensitivity parameters in the subpopulation.

UPPSALA UNIVERSITET

SV bound

For the total population, the SV bounds are

$$\mathcal{B}(\beta_R) = BF_1 \cdot BF_0 \tag{3}$$

and

$$\mathcal{B}(\beta_D) = BF_1 - P(Y=1|T=1, I_S=1)/BF_1 + P(Y=1|T=0, I_S=1)\cdot BF_0, \tag{4}$$

where $BF_1 = \frac{RR_{UY|T=1} \cdot RR_{SU|T=1}}{RR_{UY|T=1} + RR_{SU|T=1} - 1}$ and $BF_0 = \frac{RR_{UY|T=0} \cdot RR_{SU|T=0}}{RR_{UY|T=0} + RR_{SU|T=0} - 1}$.

SV bound

For the total population, the SV bounds are

$$\mathcal{B}(\beta_R) = BF_1 \cdot BF_0 \tag{3}$$

and

$$\mathcal{B}(\beta_D) = BF_1 - P(Y=1|T=1, I_S=1)/BF_1 + P(Y=1|T=0, I_S=1)\cdot BF_0, \tag{4}$$

where $BF_1 = \frac{RR_{UY|T=1} \cdot RR_{SU|T=1}}{RR_{UY|T=1} + RR_{SU|T=1} - 1}$ and $BF_0 = \frac{RR_{UY|T=0} \cdot RR_{SU|T=0}}{RR_{UY|T=0} + RR_{SU|T=0} - 1}$.

For the subpopulation, the SV bounds are

$$\mathcal{B}(\beta_{R_S}) = BF_U = \frac{RR_{UY|S=1} \cdot RR_{TU|S=1}}{RR_{UY|S=1} + RR_{TU|S=1} - 1} \tag{5}$$

and

$$\begin{aligned} \mathcal{B}(\beta_{D_S}) = \max\big[&P(Y=1|T=0, I_S=1) \cdot (BF_U - 1), \\ &P(Y=1|T=1, I_S=1) \cdot (1 - 1/BF_U)\big]. \end{aligned} \tag{6}$$

UPPSALA UNIVERSITET

Extending SV bounds to multiple selections

When extending the framework to include the selection indicator we have that the bound can be both larger and smaller than for the single selection case. This is assessed by studying the partial derivatives with respect to the selection indicator I_S of the SV bounds.

Importantly, it can be difficult for the researcher to provide plausible values for the sensitivity parameters.

Motivation for a numerical solution: R package `SelectionBias`.

SelectionBias

The function `SVboundparametersM()` calculates the sensitivity parameters and resulting bounding factors in the SV bound for the M-structure. The code, input, and output are:

```
R> SVboundparametersM(whichEst = "RR_sub",
+     Vval = matrix(c(1, 0, 0.85, 0.15), ncol = 2),
+     Uval = matrix(c(1, 0, 0.5, 0.5), ncol = 2),
+     Tcoef = c(-6.2, 1.75),
+     Ycoef = c(-5.2, 5.0, -1.0),
+     Scoef = matrix(c(1.2, 2.2, 0.0, 0.5, 2.0, -2.75, -4.0, 0.0), ncol = 4),
+     Mmodel = "L",
+     pY1_T1_S1 = 0.286,
+     pY1_T0_S1 = 0.004)
```

```
"BF_U"              1.5625
"RR_UY|S=1"         2.7089
"RR_TU|S=1"         2.3293
"Reverse treatment" TRUE
```

Zika virus example

After recoding of the treatment $\beta_R^{obs} = 0.0134$.

Assumed sensitivity parameters from the R function are applied and these values give a bound

$$\mathcal{B}(\beta_{R_S}) = \frac{RR_{UY|I_S=1} \cdot RR_{TU|I_S=1}}{RR_{UY|I_S=1} + RR_{TU|I_S=1} - 1} = \frac{2.71 \cdot 2.33}{2.71 + 2.33 - 1} = 1.56,$$

implying that: $1.56 \geq 0.0134/\beta_{R_S}$.

The causal relative risk in the subpopulation is at least $\beta_{R_S} \geq 0.0134/1.56 = 0.0086$ (those who don't have zika have a 99.14% decrease of risk of microcephaly compared to those who have zika).

UPPSALA UNIVERSITET

Assumption-free bounds

How can one bound the selection bias when no knowledge of U is available or the conditional independence assumption is not fulfilled?

Assumption-free bounds

How can one bound the selection bias when no knowledge of U is available or the conditional independence assumption is not fulfilled?

Considering the true β_R we note that the smallest value would be obtained by

$$\beta_R^{min} = \frac{\min P(Y(1)=1)}{\max P(Y(0)=1)}.$$

Decomposing and bounding $P(Y(1)=1)$ and $P(Y(0)=1)$ respectively we obtain

$$\begin{aligned}\beta_R^{min} &= \frac{P(Y(1)=1)^{min}}{P(Y(0)=1)^{max}} \\ &= \frac{P(Y=1|T=1,I_S=1)P(T=1|I_S=1)P(I_S=1)}{\min[P(T=1|I_S=1)P(I_S=1)+2P(I_S=0)+ \\ P(Y=1|T=0,I_S=1)P(T=0|I_S=1)P(I_S=1),1]}.\end{aligned}$$

Assumption-free bounds

We derive *assumption free* bounds (AF) by plugging in β_R^{min} in the bias equation, yielding a bound $\tilde{\mathcal{B}}(\beta_R)$:

$$\tilde{\mathcal{B}}(\beta_R) = \frac{\beta_R^{obs}}{\beta_R^{min}} \geq \frac{\beta_R^{obs}}{\beta_R} \tag{7}$$

and the true β_R is at least $\frac{\beta_R^{obs}}{\tilde{\mathcal{B}}(\beta_R)}$.

Similar bounds are calculated for the causal risk difference β_D and the subpopulation estimands, β_{R_S} and β_{D_S}.

SelectionBias

The function `AFbound()` takes data as input. Using the `zika_learner` data as input, the code and output are:

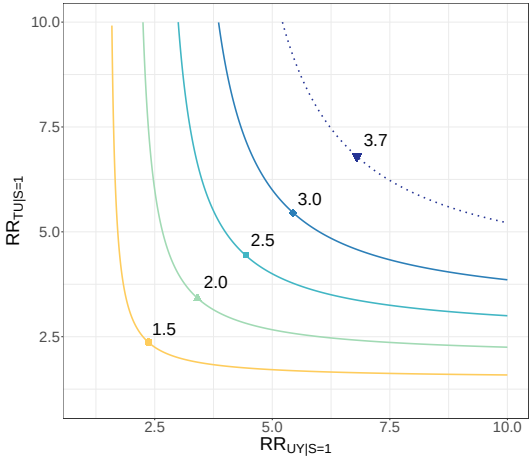
```
R> AFbound(whichEst = "RR_sub",
+     outcome = mic_ceph,
+     treatment = 1 - zika,
+     selection = sel_ind)

"AF bound" 3.5
```

From the assumption-free bound we have that the causal relative risk in the subpopulation is at least $\beta_{R_S} \geq 0.0134/3.5 = 0.0038$.

SV versus AF bound

- The SV bound is often tighter than the AF bound, especially when the treatment or outcome is rare.
- The AF bound gives an upper limit for the bounds of the selection bias.
- If $\tilde{\mathcal{B}}(\beta) < \mathcal{B}(\beta)$, then the SV bound produces values that are outside the possible range of the bias, i.e., the values are not feasible and are overly conservative.



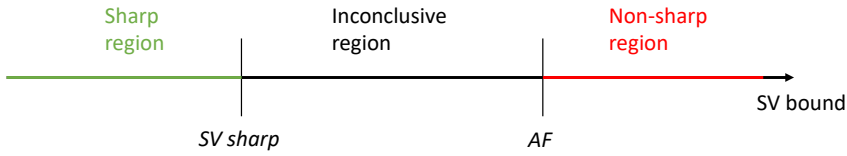
Sharpness

When are the bounds feasible?

Definition 2

A bound is sharp if the bias can be equal to the value of the bound, for an observed distribution and correctly specified sensitivity parameters.

If the bound is not sharp it may be too pessimistic.



Sharpness

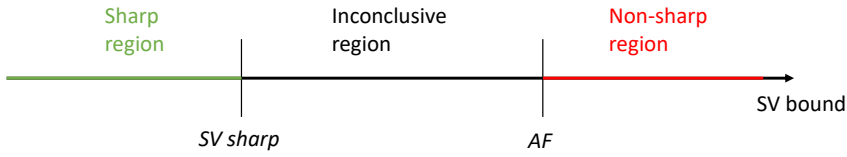
Theorem 3

Assume $\{RR_{UY|S=1}, RR_{TU|S=1}\}$ *and* $P(Y,T,U|I_S=1)$ *such that* $BF_U \leq 1/P(Y=1|T=0,I_S=1)$. *Then,* $\exists$ *a* $\beta^*_{R_S}$ *such that* $\mathcal{B}(\beta^*_{R_S}) = bias(\beta^*_{R_S})$ *and a* $\beta^*_{D_S}$ *such that* $\mathcal{B}(\beta^*_{D_S}) = bias(\beta^*_{D_S})$.

Sharpness

Theorem 3

Assume $\{RR_{UY|S=1}, RR_{TU|S=1}\}$ *and* $P(Y,T,U|I_S=1)$ *such that* $BF_U \leq 1/P(Y=1|T=0,I_S=1)$. *Then,* $\exists$ *a* $\beta^*_{R_S}$ *such that* $\mathcal{B}(\beta^*_{R_S}) = bias(\beta^*_{R_S})$ *and a* $\beta^*_{D_S}$ *such that* $\mathcal{B}(\beta^*_{D_S}) = bias(\beta^*_{D_S})$.

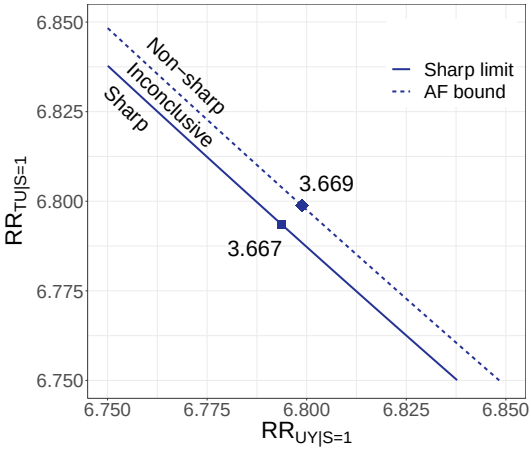


The bias can be equal to the value of the bound, in the subpopulation, when the sharp criterion is met, given that the sensitivity parameters are correct.

This can be assessed with the observed data.

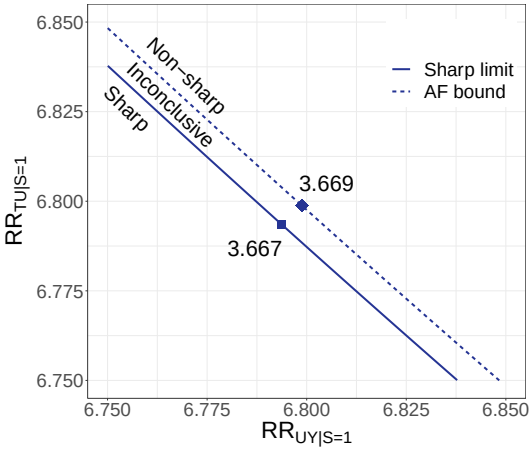
Example

The simulated data from the R package is used to illustrate the sharpness.



Example

The simulated data from the R package is used to illustrate the sharpness.



- Dotted curve: the SV bound is equal to the AF bound$= 3.669$.
- Solid curve: sharp limit$= 3.667$.
- The sharp limit is almost identical to the AF bound. This is because $P(T = 1|I_S = 1) \approx 1$.

UPPSALA UNIVERSITET

Total population

- There is no corresponding result for sharp bounds for the total population.
  - This comes from the construction of the bounds.
  - $bias(\beta_R) = \frac{\beta_R^{obs}}{\beta_R} \leq \frac{P(Y=1|T=1,I_S=1)}{P(Y=1|T=0,I_S=1)} \Big/ \frac{\min_s P(Y=1|T=1,I_S=s)}{\max_s P(Y=1|T=0,I_S=s)}$

    $= \frac{P(Y=1|T=1,I_S=1)}{\min_s P(Y=1|T=1,I_S=s)} \cdot \frac{\max_s P(Y=1|T=0,I_S=s)}{P(Y=1|T=0,I_S=1)} \leq BF_1 \cdot BF_0.$

Total population

- There is no corresponding result for sharp bounds for the total population.
  - This comes from the construction of the bounds.
  - $bias(\beta_R) = \frac{\beta_R^{obs}}{\beta_R} \leq \frac{P(Y=1|T=1,I_S=1)}{P(Y=1|T=0,I_S=1)} \Big/ \frac{\min_s P(Y=1|T=1,I_S=s)}{\max_s P(Y=1|T=0,I_S=s)}$

    $= \frac{P(Y=1|T=1,I_S=1)}{\min_s P(Y=1|T=1,I_S=s)} \cdot \frac{\max_s P(Y=1|T=0,I_S=s)}{P(Y=1|T=0,I_S=1)} \leq BF_1 \cdot BF_0.$
- If $P(Y=1|T=t,I_S=1) \neq P(Y=1|T=t,I_S=0)$, $t=0,1$, the first inequality is strict $\Rightarrow$ the bias cannot be as large as the bound.
- The bias is not greater than the SV bound.
- Corresponding results for the risk difference.

UPPSALA UNIVERSITET

SelectionBias

Sharpness can be assessed in the R package. The code, input, and output are:

```
R> SVboundsharp(BF_U = 1.56,
+     pY1_T0_S1 = 0.286,
+     SVbound = 1.56,
+     AFbound = 3.5)

"SV bound is sharp."
```

UPPSALA UNIVERSITET

Conclusions

- Study population inclusion/exclusion criteria can result in selection bias.
- Sensitivity analysis can help to assess the magnitude of selection bias.
- SV bounds are extended to multiple selections. Results on variation independence and conditions for sharp bounds.
- Assumption free bounds and R package `SelectionBias`.
- Results are applied in the tutorial `zika_learner` and in a study of the effect of pre-term birth on type 1 diabetes.

UPPSALA UNIVERSITET

- Model and selection bias
- SV bound
- AF bound
- Comparisons
- Conclusions

References

de Araújo, T. V. B., de Alencar Ximenes, R. A., de Barros Miranda-Filho, D., Souza, W. V., Montarroyos, U. R., de Melo, A. P. L., ... Oliveira, V. F. (2018). Association between microcephaly, Zika virus infection, and other risk factors in Brazil: final report of a case-control study. *The Lancet infectious diseases*, 18(3), 328-336.

Sjölander, A. (2020). A note on a sensitivity analysis for unmeasured confounding, and the related E-value. *Journal of Causal Inference*, 8 (1), 229–248.

Smith, L. H. and T. J. VanderWeele (2019). Bounding bias due to selection. *Epidemiology*, 30 (4), 509–516.

Zetterstrom, S. and Waernbaum, I. (2022). Selection bias and multiple inclusion criteria in observational studies. *Epidemiologic methods*, 11(1).

Zetterstrom, S. and Waernbaum, I. (2023). `Selection bias:` an R package for bounding selection bias. (arxiv.org/abs/2302.06518)