UPPSALA UNIVERSITET

- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work

# Bounds for selection bias using outcome probabilities

Stina Zetterström

Department of Statistics, Uppsala University

2025-03-27

UPPSALA UNIVERSITET


- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work


# 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:
  - unmeasured confounding
  - selection bias
- Selection bias can arise from missing data or when the study population is constructed, often by inclusion or exclusion criteria.

# "Data and study population"

UPPSALA UNIVERSITET

- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work

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.

Crump et al (2023) in BMJ.

# "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.

Crump et al (2023) in BMJ.

We are interested in bounding a possible bias from these selections.

UPPSALA UNIVERSITET



# Difference of the bounds

Different bounds useful depending on the situation and prior knowledge.

| Bound | Includes unknown sensitivity parameters? | Includes data? | Relies on additional assumptions? | Sensitivity parameters |
|---|---|---|---|---|
| SV | ✓ | | ✓ | Ratios of probabilities |
| AF | | ✓ | | None |
| GAF | ✓ | ✓ | ✓ | Probabilities |
| CAF | ✓ | ✓ | | Counterfactual probabilities |
| Sharp bounds | ✓ | ✓ | ✓ | Ratios of probabilities |

UPPSALA UNIVERSITET



# Difference of the bounds

Different bounds useful depending on the situation and prior knowledge.

| Bound | Includes unknown sensitivity parameters? | Includes data? | Relies on additional assumptions? | Sensitivity parameters |
|---|---|---|---|---|
| SV | ✓ | | ✓ | Ratios of probabilities |
| AF | | ✓ | | None |
| GAF | ✓ | ✓ | ✓ | Probabilities |
| CAF | ✓ | ✓ | | Counterfactual probabilities |
| Sharp bounds | ✓ | ✓ | ✓ | Ratios of probabilities |

We focus on the risk ratio in the total population, but corresponding results for the selected population and risk difference are presented in the paper.

UPPSALA UNIVERSITET



# Outline

Model and selection bias

Bounds

Comparative study

Conclusion and future work

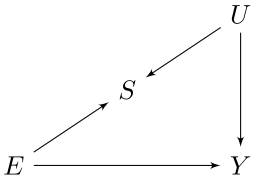
UPPSALA UNIVERSITET

- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work

# Model and notation

Variables in the model:

- Binary exposure variable, $E$.
- Binary potential outcomes, $Y_e,\ e = 0, 1$.
- Selection variable, $S$.
- Vector of unmeasured variables, $U$.
- Vector of observed baseline covariates, $X$.




*Example structure.*

# Model and notation

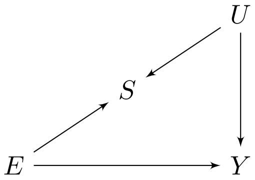
Variables in the model:

- Binary exposure variable, $E$.
- Binary potential outcomes, $Y_e,\ e = 0, 1$.
- Selection variable, $S$.
- Vector of unmeasured variables, $U$.
- Vector of observed baseline covariates, $X$.

Assumptions:

- Consistency, $Y = E \cdot Y_1 + (1 - E) \cdot Y_0$.
- Conditional exchangeability, $Y_e \perp\!\!\!\perp E | X,\ e = 0, 1$.
- $Y_e \not\perp\!\!\!\perp E | (S = 1, X),\ e = 0, 1$.
- All analysis is done conditional on $X = x$.
- Ignore sampling variability $\rightarrow$ the observed means are treated as an approximation of the corresponding asymptotic mean:

$$\frac{1}{n} \sum_{i:E=e,S=1} Y_i \xrightarrow{p} p(Y = 1 | E = e, S = 1).$$




*Example structure.*

# Risk ratio and selecion bias

The method applies to the risk ratio and risk difference in the total and selected population.

Focus here: causal risk ratio in the total population, defined as

$$RR_T = \frac{P(Y_1 = 1)}{P(Y_0 = 1)}.$$

# Risk ratio and selecion bias

The method applies to the risk ratio and risk difference in the total and selected population.

Focus here: causal risk ratio in the total population, defined as

$$RR_T = \frac{P(Y_1 = 1)}{P(Y_0 = 1)}.$$

Under selection, $S = 1$ we define the observed risk ratio $RR^{obs}$ as

$$RR^{obs} = \frac{P(Y = 1|E = 1, S = 1)}{P(Y = 1|E = 0, S = 1)}.$$

# Risk ratio and selecion bias

The method applies to the risk ratio and risk difference in the total and selected population.

Focus here: causal risk ratio in the total population, defined as

$$RR_T = \frac{P(Y_1 = 1)}{P(Y_0 = 1)}.$$

Under selection, $S = 1$ we define the observed risk ratio $RR^{obs}$ as

$$RR^{obs} = \frac{P(Y = 1|E = 1, S = 1)}{P(Y = 1|E = 0, S = 1)}.$$

The selection bias is defined as a ratio of the risk ratios

$$Bias(RR_T) = \frac{RR^{obs}}{RR_T} = \frac{\frac{P(Y=1|E=1,S=1)}{P(Y=1|E=0,S=1)}}{\frac{P(Y_1=1)}{P(Y_0=1)}}.$$

UPPSALA
UNIVERSITET


- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work


# Outline


Model and selection bias

Bounds

Comparative study

Conclusion and future work

UPPSALA UNIVERSITET

- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work

# Potential outcome probabilities

The bounds are constructed by bounding each potential outcome probability using both data and sensitivity parameters.

The potential outcome probabilities can be decomposed as

$$\begin{aligned} P(Y_e = 1) = {} & P(Y = 1|E = e, S = 1)P(S = 1|E = e) \\ & + P(Y = 1|E = e, S = 0)P(S = 0|E = e), \quad e = 0, 1. \end{aligned}$$

UPPSALA UNIVERSITET


- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work


# Potential outcome probabilities

The bounds are constructed by bounding each potential outcome probability using both data and sensitivity parameters.

The potential outcome probabilities can be decomposed as

$$\begin{aligned} P(Y_e = 1) = {} & P(Y = 1|E = e, S = 1)P(S = 1|E = e) \\ & + P(Y = 1|E = e, S = 0)P(S = 0|E = e), \quad e = 0, 1. \end{aligned}$$

$P(S = 1|E = e) \geq P(E = e|S = 1)P(S = 1)$ if the proportion of the selected subjects is known.

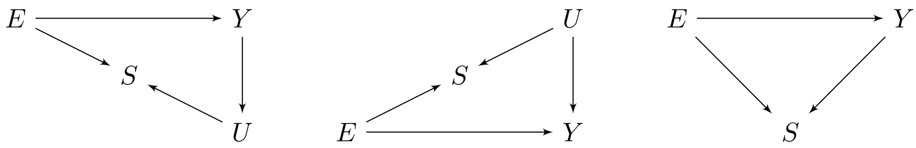
Only the probability $P(Y = 1|E = e, S = 0)$ is unobserved. This can be bounded under a conditional independence assumption.

# Conditional independence assumption

## Assumption 1

*There exists an unmeasured variable(s) U such that* $Y \perp\!\!\!\perp S | E, U$.

There are several structures for which this property holds, (a) and (b), but also structures such that it is not fulfilled, (c):




(a) (b) (c)

# Sensitivity parameters

The presence of a valid $U$, according to Assumption 1, implies that

$$\begin{aligned}
&\min_{e,u} P(Y=1|E=e,U=u) \\
&< P(Y=1|E=e,S=0) \\
&< \max_{e,u} P(Y=1|E=e,U=u).
\end{aligned}$$

# Sensitivity parameters

The presence of a valid $U$, according to Assumption 1, implies that

$$\begin{aligned}
&\min_{e,u} P(Y=1|E=e,U=u) \\
&< P(Y=1|E=e,S=0) \\
&< \max_{e,u} P(Y=1|E=e,U=u).
\end{aligned}$$

The sensitivity parameters are defined as

$$m_T = \min_{e,u} P(Y=1|E=e,U=u)$$

and

$$M_T = \max_{e,u} P(Y=1|E=e,U=u).$$

UPPSALA UNIVERSITET

- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work

# Sensitivity parameters

The presence of a valid $U$, according to Assumption 1, implies that

$$\begin{aligned}&\min_{e,u} P(Y=1|E=e,U=u)\\ &< P(Y=1|E=e,S=0)\\ &< \max_{e,u} P(Y=1|E=e,U=u).\end{aligned}$$

The sensitivity parameters are defined as

$$m_T = \min_{e,u} P(Y=1|E=e,U=u)$$

and

$$M_T = \max_{e,u} P(Y=1|E=e,U=u).$$

Observe that one unknown probability is replaced by another unknown probability. This only makes sense if $P(Y=1|E=e,U=u)$ are easier to guess.

# Generalized assumption-free (GAF) bounds

Combining the probabilities observed from data with the sensitivity parameters results in bounds for the relative risk, $RR_T$:

$$LB_T < RR_T < UB_T \tag{1}$$

with the lower bound defined as

$$LB_T = \frac{P(Y=1, E=1, S=1) + [1 - P(E=1, S=1)] \cdot m_T}{P(Y=1, E=0, S=1) + [1 - P(E=0, S=1)] \cdot M_T}$$

and the upper bound

$$UB_T = \frac{P(Y=1, E=1, S=1) + [1 - P(E=1, S=1)] \cdot M_T}{P(Y=1, E=0, S=1) + [1 - P(E=0, S=1)] \cdot m_T}.$$

The GAF bounds are equal to the AF bounds when $m_T = 0$ and $M_T = 1$.

# Properties of the GAF bounds

**Feasible region:**

- The sensitivity parameters are probabilities $\Rightarrow$ restricted by 0 and 1.

UPPSALA UNIVERSITET


- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work


# Properties of the GAF bounds

**Feasible region:**

- The sensitivity parameters are probabilities $\Rightarrow$ restricted by 0 and 1.
- From construction: $m_T < P(Y = 1|E = e, S = 1) < M_T \Rightarrow$
  - $0 \leq m_T < \min_e P(Y = 1|E = e, S = 1)$
  - $\max_e P(Y = 1|E = e, S = 1) < M_T \leq 1$

UPPSALA UNIVERSITET


- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work


# Properties of the GAF bounds

**Feasible region:**

- The sensitivity parameters are probabilities $\Rightarrow$ restricted by 0 and 1.
- From construction: $m_T < P(Y=1|E=e,S=1) < M_T \Rightarrow$
  - $0 \leq m_T < \min_e P(Y=1|E=e,S=1)$
  - $\max_e P(Y=1|E=e,S=1) < M_T \leq 1$
- GAF bounds always cover the null effect.
  - $LB_T < 1 < UB_T$

# Properties of the GAF bounds

**Feasible region:**

- The sensitivity parameters are probabilities $\Rightarrow$ restricted by 0 and 1.
- From construction: $m_T < P(Y = 1|E = e, S = 1) < M_T \Rightarrow$
  - $0 \leq m_T < \min_e P(Y = 1|E = e, S = 1)$
  - $\max_e P(Y = 1|E = e, S = 1) < M_T \leq 1$
- GAF bounds always cover the null effect.
  - $LB_T < 1 < UB_T$

**Sharpness:**

- A bound is sharp if it can be equal to the causal estimand.

UPPSALA UNIVERSITET



# Properties of the GAF bounds

**Feasible region:**

- The sensitivity parameters are probabilities $\Rightarrow$ restricted by 0 and 1.
- From construction: $m_T < P(Y = 1|E = e, S = 1) < M_T \Rightarrow$
  - $0 \leq m_T < \min_e P(Y = 1|E = e, S = 1)$
  - $\max_e P(Y = 1|E = e, S = 1) < M_T \leq 1$
- GAF bounds always cover the null effect.
  - $LB_T < 1 < UB_T$

**Sharpness:**

- A bound is sharp if it can be equal to the causal estimand.
- In the GAF bounds in the total population, both $P(E = 1) = 1$ and $P(E = 0) = 1$, in order to reduce the number of guesses. However, this is logically impossible $\Rightarrow$ GAF bounds are not sharp.

UPPSALA UNIVERSITET


- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work


# Outline


Model and selection bias

Bounds

Comparative study

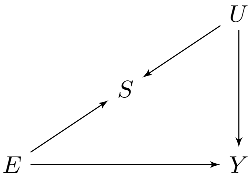
Conclusion and future work

# Comparative study setup

The GAF and the AF bounds are compared to Smith and VanderWeele's (SV) bounds in a numerical example.

The model is parameterized as

- $p(U=1) = \text{expit}(\theta_1)$
- $p(E=1) = \text{expit}(\theta_2)$
- $p(S=1 \mid E, U) = \text{expit}(\alpha + \beta E + \gamma U)$
- $p(Y=1 \mid E, U) = \text{expit}(\delta + \lambda E + \psi U)$




The coefficients $\beta$, $\gamma$, $\lambda$, and $\psi$ are independently drawn from $N(0,1)$.

The parameters $\theta_1$, $\theta_2$, $\alpha$ and $\delta$ are set to obtain different marginal probabilities.

UPPSALA UNIVERSITET

- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work

# Simulation setup

1000 distributions are generated for each combination, but only $\sim$500 are used. SV's bounds require the observed risk ratio to be larger than the causal risk ratio, so only these distributions are used and comparisons are only made for lower bounds.

UPPSALA UNIVERSITET

- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work

# Simulation setup

1000 distributions are generated for each combination, but only ∼500 are used. SV's bounds require the observed risk ratio to be larger than the causal risk ratio, so only these distributions are used and comparisons are only made for lower bounds.

Two measures of the performance of the bounds:

1. Distance between the causal estimand and the bounds measured on the same scale as the estimand:
   - $\Delta_{bound} = |\log RR - \log bound|$
2. The proportions of distributions when the SV bounds are tighter than the GAF and AF bounds, $p_{bound}$.

# Simulation setup

1000 distributions are generated for each combination, but only $\sim$500 are used. SV's bounds require the observed risk ratio to be larger than the causal risk ratio, so only these distributions are used and comparisons are only made for lower bounds.

Two measures of the performance of the bounds:

1. Distance between the causal estimand and the bounds measured on the same scale as the estimand:
   - $\Delta_{bound} = |\log RR - \log bound|$
2. The proportions of distributions when the SV bounds are tighter than the GAF and AF bounds, $p_{bound}$.

UPPSALA UNIVERSITET

- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work

# Simulation results $RR_T$

| $P(U=1)$ | $P(E=1)$ | $P(Y=1)$ | $P(S=1)$ | $p_{GAF}$ | $p_{AF}$ | $\Delta_{GAF}$ | $\Delta_{AF}$ | $\Delta_{SV}$ | $\log RR_T$ |
|---|---|---|---|---|---|---|---|---|---|
| 0.20 | 0.05 | 0.05 | 0.50 | 0.92 | 1.00 | 1.31 | 6.22 | 0.32 | 0.00 |
| 0.20 | 0.05 | 0.05 | 0.80 | 0.83 | 1.00 | 1.00 | 5.04 | 0.29 | −0.07 |
| 0.20 | 0.05 | 0.20 | 0.50 | 0.91 | 1.00 | 0.94 | 4.94 | 0.25 | −0.05 |
| 0.20 | 0.05 | 0.20 | 0.80 | 0.83 | 1.00 | 0.74 | 3.96 | 0.25 | −0.14 |
| 0.20 | 0.20 | 0.05 | 0.50 | 0.89 | 1.00 | 1.11 | 4.96 | 0.30 | −0.01 |
| 0.20 | 0.20 | 0.05 | 0.80 | 0.86 | 1.00 | 0.96 | 3.99 | 0.28 | 0.02 |
| 0.20 | 0.20 | 0.20 | 0.50 | 0.90 | 1.00 | 0.95 | 3.63 | 0.27 | −0.03 |
| 0.20 | 0.20 | 0.20 | 0.80 | 0.84 | 1.00 | 0.80 | 2.79 | 0.26 | −0.01 |
| 0.50 | 0.05 | 0.05 | 0.50 | 0.89 | 1.00 | 1.21 | 6.19 | 0.33 | −0.03 |
| 0.50 | 0.05 | 0.05 | 0.80 | 0.84 | 1.00 | 0.94 | 4.99 | 0.30 | −0.10 |
| 0.50 | 0.05 | 0.20 | 0.50 | 0.90 | 1.00 | 0.99 | 4.93 | 0.28 | −0.03 |
| 0.50 | 0.05 | 0.20 | 0.80 | 0.86 | 1.00 | 0.84 | 3.96 | 0.28 | −0.04 |
| 0.50 | 0.20 | 0.05 | 0.50 | 0.91 | 1.00 | 1.15 | 4.99 | 0.33 | −0.02 |
| 0.50 | 0.20 | 0.05 | 0.80 | 0.84 | 1.00 | 0.87 | 3.95 | 0.31 | −0.06 |
| 0.50 | 0.20 | 0.20 | 0.50 | 0.89 | 1.00 | 0.93 | 3.61 | 0.28 | −0.09 |
| 0.50 | 0.20 | 0.20 | 0.80 | 0.83 | 1.00 | 0.77 | 2.78 | 0.27 | −0.04 |

# Outline



Model and selection bias

Bounds

Comparative study

Conclusion and future work

# Conclusions and future work

- Study population inclusion/exclusion criteria can result in selection bias.
- Sensitivity analysis can help to assess the magnitude of selection bias.
- Different types of bounds are useful in different settings.
- GAF bounds can have more intuitive sensitivity parameters compared to other bounds based on relative risks but can be conservative.
- GAF bounds is tighter than SV in some settings, especially when $P(S = 1)$ is higher.

# Conclusions and future work

- Study population inclusion/exclusion criteria can result in selection bias.
- Sensitivity analysis can help to assess the magnitude of selection bias.
- Different types of bounds are useful in different settings.
- GAF bounds can have more intuitive sensitivity parameters compared to other bounds based on relative risks but can be conservative.
- GAF bounds is tighter than SV in some settings, especially when $P(S = 1)$ is higher.
- Bounds are defined conditional on the covariates.
- Sampling variability not considered.

# References

Crump, C., Sundquist, J., McLaughlin, M. A., Dolan, S. M., Govindarajulu, U., Sieh, W., and Sundquist, K. (2023). Adverse pregnancy outcomes and long term risk of ischemic heart disease in mothers: national cohort and co-sibling study. BMJ, 380:e072112.

Peña, J. (2022). Simple yet sharp sensitivity analysis for unmeasured confounding. *Journal of Causal Inference*, 10(1), 1-17.

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.

Waernbaum, I., Dahlquist, G. and Lind, T (2019). Perinatal risk factors for type 1 diabetes revisited: a population-based register study. *Diabetologia* 62, 1173–1184.

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

UPPSALA UNIVERSITET

UPPSALA UNIVERSITET



# Outline


Model and selection bias

Bounds

Comparative study

Conclusion and future work

UPPSALA UNIVERSITET

- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work

# Preterm birth and type 1 diabetes

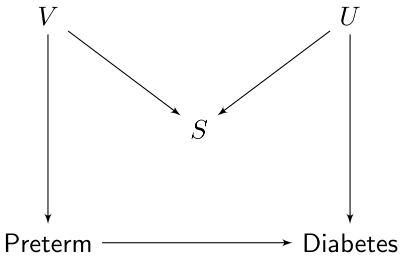
A case-control study by Waernbaum, Dahlquist and Lind (2019) investigated the causal effect of preterm birth ($E$) on type 1 diabetes ($Y$).

Three restrictions on the study population were made:

- Nordic mothers
- Singleton births
- Non-diabetic mothers

These comprise the selection variable, $S$.

$Y_e \perp\!\!\!\perp E|(S=1, U=u)$, for $e=0,1$ can be assumed to hold.



# Preterm birth and type 1 diabetes

The exposure probabilities are known from the data, but the outcome probabilities are not known since this is a case-control study. However, for the sake of illustration, values are assumed. The probabilities are:

- $P(E = 1|S = 1) = 0.005$
- $P(E = 0|S = 1) = 0.995$
- $P(Y = 1|E = 1, S = 1) = 0.00013$
- $P(Y = 1|E = 0, S = 1) = 0.00025$

# Preterm birth and type 1 diabetes

The GAF bounds are

$$LB_S = \frac{0.00013 \cdot 0.005 + 0.995 \cdot m_S}{0.00025 \cdot 0.995 + 0.005 \cdot M_S}$$

and

$$UB_S = \frac{0.00013 \cdot 0.005 + 0.995 \cdot M_S}{0.00025 \cdot 0.995 + 0.005 \cdot m_S}.$$

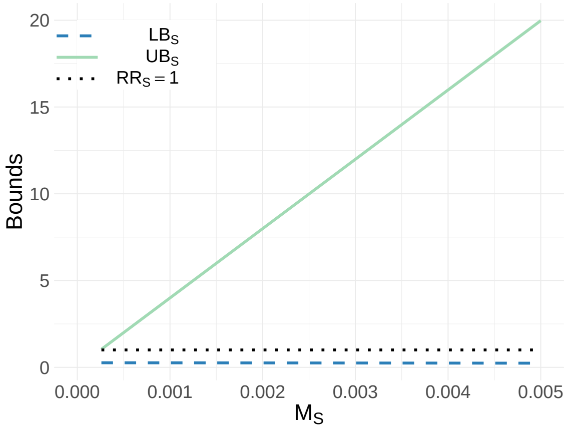
The maximum value of $m_S$ is very small $\Rightarrow UB_S$ is dominated by $M_S$.

$M_S$ is varied and $m_S = 0.000065$.

UPPSALA UNIVERSITET

- Model and selection bias
- Bounds
- Comparative study
- Conclusion and future work

# Preterm birth and type 1 diabetes




$$RR^{obs} = 0.53$$