PHS2000B Lab 2

Sensitivity Analysis

01/27/2020

Contents

1 Background

1.1 The problem

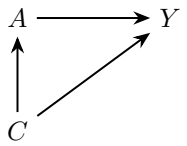
As we saw in Lab 1, inferring whether an observed association represents a true causal effect requires additional assumptions beyond those typically required for statistical modeling. Chief amongst these is the assumption that the exposure/treatment groups are exchangeable, i.e. that the groups each represent the counterfactual experience of the other if their exposure/treatment status were reversed. More formally we said that exchangeability implied a statistical independence between the potential outcomes Y_a and the exposure/treatment received, i.e.

$$Y_a \perp\!\!\!\perp A$$

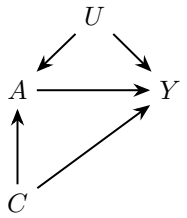
Connecting this idea to DAGs, we also saw that exchangeability could be determined graphically via the backdoor criterion: variable A is exchangeable with respect descedant Y if there are no back door paths connecting A to Y on their causal graph. Under randomization, exchangeability is ensured by design because the determination of who gets treated and who doesn't is solely determined by random chance. Graphically we can see this because the randomization of A implies that there are no arrows into A as nothing causes A other than a proverbial flip of the coin.

$$A \longrightarrow Y$$

In the observational setting, we often must make strong assumptions that treatment or exposure groups are conditionally exchangeable within levels of a sufficient set of covariates C, i.e. $Y_a \perp\!\!\!\perp A \mid C$, where C is set that covers all possible backdoor paths between A and Y.



In practice, unless we know the precise mechanism whereby certain people come to be exposed/treated while others don't, we often can't be certain that we've measured all relevant covariates (or for that matter that we've measured them without error and have modeled the correct functional form of their relationship to the treatment variable). In this case there may still be other unobserved covariates U representing additional paths not closed by conditioning on C.



These remaining backdoor paths through U mean that any estimate of the association between A on Y from the observed data is likely to be a biased estimate of the true relationship between A and Y. With no data on U and perhaps little information about what lurking variables U may contain, at this point it may seem as if all were lost, however it turns out that we can develop reasonable bounds on the amount of bias that U must produce to explain away an observed result and then we can apply our expert subject matter knowledge to determine how reasonable it is that such a U might exist. This is the concept behind sensitivity analysis, the idea that one can investigate how conclusions might change under different hypothetical assumptions about the nature of the data generating process.

1.2 Cornfield's conditions

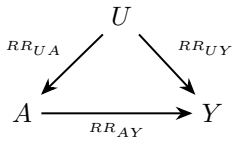
In Epi 201, you developed an intuition about the bounds of bias due to confounding through the Cornfield conditions. Originally conceived as a means of rebutting Fisher's hypothesis that there was a "smoking gene" that both made people more likely to smoke and caused them to develop lung cancer, the conditions stated that in order for a single confounder to explain away the observed association the association between the confounder and the exposure must be at least as large as the association between the exposure and outcome, i.e.

$$RR_{UA} \geq RR_{AY}$$

and the association between the confounder and outcome must equally be at least as large as the association between exposure and outcome, i.e.

$$RR_{UY} \geq RR_{AY}$$

These conditions applied to the restricted case in which there was a single unobserved confounder with no interaction between exposure and confounder and the exposure, confounder, and outcome were all binary.



However, as mentioned in class it is possible to generate more nuanced bounds that make no assumptions about the structure of confouding bias and the nature of the variables under consideration. We develop these further in the next section.

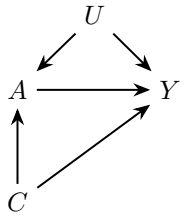
2 Sensitivity analysis without assumptions

In a seminal paper in the 1970s Ollie Miettinen had the insight that the observed risk ratio can, in a sense, be decomposed into a component representing the true causal risk ratio and a *bias factor* representing how much the observed over or under shoots this mark.

$$RR_{obs} = BF \cdot RR_{true} \tag{1}$$

For example if we estimated an observed risk ratio of 3 this could reflect a true causal ratio of 1 and a bias factor of 3 (i.e. no true effect, the observed association is totally explained by confounding), a true causal risk ratio of 3 and a bias factor of 3 (i.e. no residual confounding so association is causation), or something in between like a true causal risk ratio of 2 and a bias factor of 1.5.

Returning to the common scenario in observational research where we have adjusted for a set of measured covariates C but are concerned that there may be remaining U the might bias the result, we can use the counterfactual logic we learned in Lab 1 to develop expressions for RR_{true} and RR_{obs} then relate it to BF using equation 1.



First it should be clear that RR_{obs} in this case is just the adjusted risk ratio from our observational study, i.e.

$$RR_{obs,c} = \frac{P(Y = 1 \mid A = 1, C = c)}{P(Y = 1 \mid A = 0, C = c)}$$

Then using the rules of d-separation on the graph above, we know that the effect of A on Y is identified if and only if we condition on both C and U, thus

$$RR_{true,c,u} = \frac{P(Y_1 \mid C = c, U = u)}{P(Y_0 \mid C = c, U = u)} = \frac{P(Y = 1 \mid A = 1, C = c, U = u)}{P(Y = 1 \mid A = 0, C = c, U = u)}$$

because

$$Y_a \perp\!\!\!\perp A \mid (C, U)$$

assuming consistency and positivity. We want to relate $RR_{true,c,u}$ and $RR_{obs,c}$ to develop expressions for the bias factor BF, however we can't yet as the former is conditional on U and C while the latter is only

conditional on C. Using the basic rules of probability we can get $RR_{true,c}$, i.e. the true risk ratio conditional on C alone, but standardizing (taking the weighted sum over all values of U).

$$RR_{true,c} = \frac{\sum_u P(Y = 1 \mid A = 1, C = c, U = u)P(U = u \mid C = c)}{\sum_u P(Y = 1 \mid A = 0, C = c, U = u)P(U = u \mid C = c)}$$

Therefore we have shown that the bias factor is equal to

$$BF = \frac{RR_{obs,c}}{RR_{true,c}} = \frac{\frac{P(Y=1|A=1,C=c)}{P(Y=1|A=0,C=c)}}{\frac{\sum_u P(Y=1|A=1,C=c,U=u)P(U=u|C=c)}{\sum_u P(Y=1|A=0,C=c,U=u)P(U=u|C=c)}}$$

In their seminal paper, Ding and Vanderweele show that bounds for this bias factor can be developed without any further assumptions about Y, A, C, or U.

$$BF \leq \frac{RR_{UA} \times RR_{UY}}{RR_{UA} + RR_{UY} - 1} \tag{2}$$

That is the largest that this BF could be can be represented by just two parameters RR_{UY} and RR_{UA}. Where RR_{UY} is the largest possible risk ratio for the outcome comparing any two values of the unmeasured confounder within either stratum of the exposure

$$RR_{UY} = \max\left(\frac{\max_u P(Y = 1|A = 0, c, u)}{\min_u P(Y = 1|A = 0, c, u)}, \frac{\max_u P(Y = 1|A = 1, c, u)}{\min_u P(Y = 1|A = 1, c, u)}\right)$$

and RR_{UA} is the maximum risk ratio for a single value of the unmeasured confounder comparing the two exposure levels.

$$RR_{UA} = \max_u \frac{P(u|A = 1, c)}{P(u|A = 0, c)}$$

Thus a general road map for developing bounds on confounding bias in most observational settings is: 4

1. Calculate the adjusted risk ratio $RR_{obs,c}$ relating the exposure to the outcome after conditioning on all measured confounders C.
2. Using your subject matter expertise, come up with suitable estimates of RR_{UY} and RR_{UA} for the most likely unmeasured confouders in U. Alternatively, if not much is known about the structure of U explore a range of values.
3. Using equation 2 calculate the largest bias factor that could results from these values of RR_{UY} and RR_{UA}.
4. Using equation 1 relate this to potential values that the true risk ratio could conceivably take on $RR_{true,c}$.

2.1 Examples

You conduct an observational study in which you estimate an adjusted risk ratio of 1.8 (95% CI: 1.4 to 2.2) for the assocation between an exposure and outcome conditional on a set of covariates. You're concerned that there was another variable Z that could confound the association you observed, you would have really liked to have measured Z but alas it just wasn't possible given the constraints of your study. A recent meta-analysis of the effect of Z on your outcome suggests the Z increases the risk of the outcome by a factor of 2.3.

How large must the association between Z and your exposure of interest be for Z to completely explain away the observed association?

For confouding to completely explain away the observed association would imply that the bias factor equals the observed risk ratio, i.e. $BF = 1.8$. Applying equation 1 with the given value of $RR_{UY} = 2.3$ then

$$\begin{aligned} BF &\leq \frac{RR_{UA} \times RR_{UY}}{RR_{UA} + RR_{UY} - 1} \\ 1.8 &\leq \frac{2.3RR_{UA}}{RR_{UA} + 2.3 - 1} \\ 1.8 &\leq \frac{2.3RR_{UA}}{RR_{UA} + 1.3} \\ 1.8(RR_{UA} + 1.3) &\leq 2.3RR_{UA} \\ 1.8RR_{UA} + 2.34 &\leq 2.3RR_{UA} \\ 2.34 &\leq 0.5RR_{UA} \\ RR_{UA} &\geq 4.68 \end{aligned}$$

Therefore in order to completely explain away the effect RR_{UA} must be at least 4.68.

How large must the association between Z and your exposure of interest be for Z to shift the observed 95% CI to include the null?

Applying the same logic as before, except that now the bias factor only needs to be as large as the 95% CI limit closest to the null.

$$\begin{aligned} BF &\leq \frac{RR_{UA} \times RR_{UY}}{RR_{UA} + RR_{UY} - 1} \\ 1.4 &\leq \frac{2.3RR_{UA}}{RR_{UA} + 2.3 - 1} \\ 1.4 &\leq \frac{2.3RR_{UA}}{RR_{UA} + 1.3} \\ 1.4(RR_{UA} + 1.3) &\leq 2.3RR_{UA} \\ 1.4RR_{UA} + 1.82 &\leq 2.3RR_{UA} \\ 1.82 &\leq 0.9RR_{UA} \\ RR_{UA} &\geq 2.02 \end{aligned}$$

Therefore in order to shift the 95% CI to include the null the relationship between Z and A only needs to be at least 2.02.

3 E-value

Alternatively, what if one simply wished to know how much confouding is necessary to explain away the observed association? That is what would be the values of RR_{UY} and RR_{UA} necessary for $RR_{true,c} = 1$? Using equations 1 and 2 again, we can see that if we assume a value for one or the other we can determine the value of its counterpart when $RR_{true,c} = 1$. What if we just wanted to know in the special case that they are equivalent, i.e. $RR_{eq} = RR_{UY} = RR_{UA}$. Then the bias factor simplifies to

$$BF \leq \frac{RR_{eq}^2}{2RR_{eq} - 1}$$

And then relating that to equation 1 and doing some simple algebra we can show (see section 7.1 if you want to see full proof)

$$\text{E-value} = RR_{eq} = RR_{obs} + \sqrt{RR_{obs}(RR_{obs} - 1)} \tag{3}$$

Thus RR_{eq} can be calculated just using the observed risk ratio. The quantity RR_{eq} is therefore given a special name that you may be familiar with: we call it the E-value. It represents *the minimum strength of association on the risk ratio scale that an unmeasured confounder would need to have with both the exposure and the outcome, conditional on the measured covariates, to fully explain away a specific exposure-outcome relationship.*

3.1 Examples

Using the same example as above:

> Provide a measure of the evidence for causality (E-value) for both the point estimate and confidence interval. Interpret these quantities.

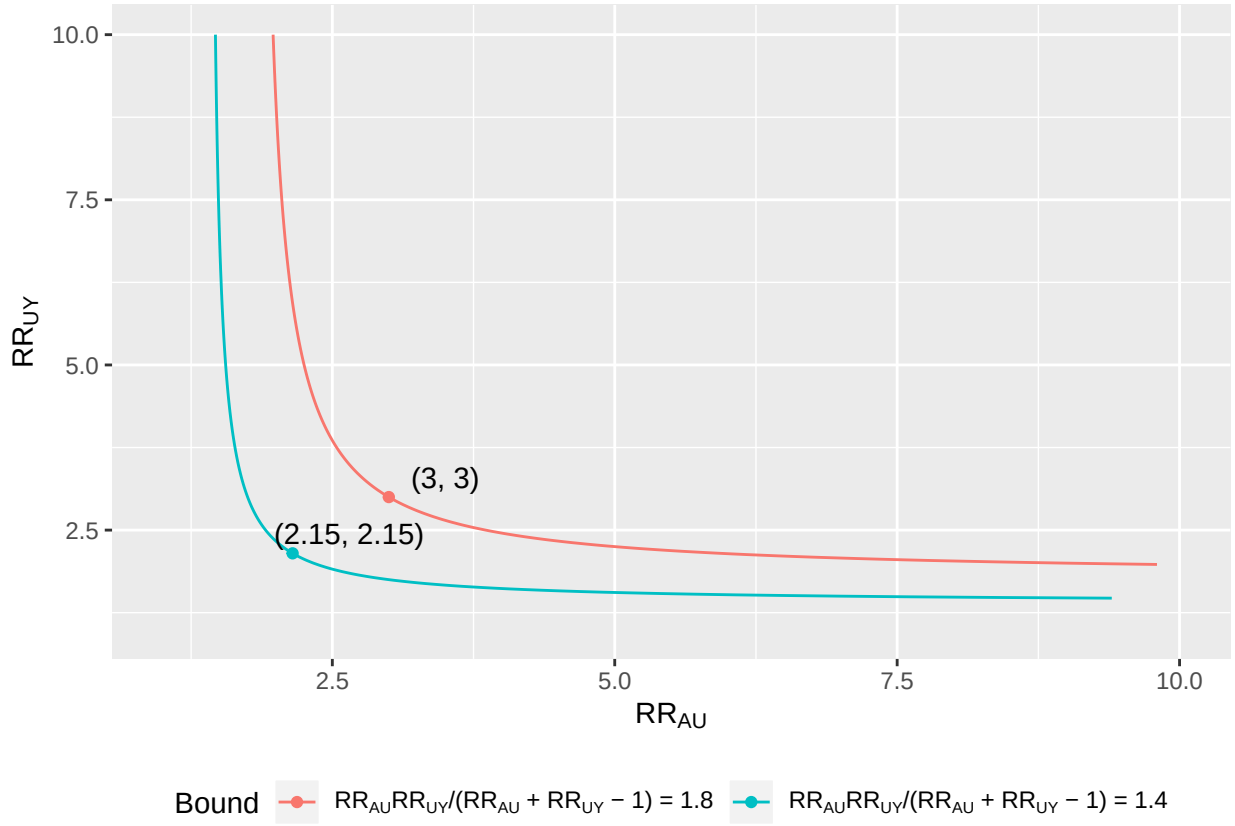
> The observed RR is 1.8 and it's lower CI limit on the RR scale is 1.4.
>
> $$\begin{aligned} \text{E-value} &= RR + \sqrt{RR(RR-1)} \\ &= 1.8 + \sqrt{1.8(1.8-1)} \\ &= 3 \end{aligned}$$
>
> $$\begin{aligned} \text{E-value} &= RR + \sqrt{RR(RR-1)} \\ &= 1.4 + \sqrt{1.4(1.4-1)} \\ &= 2.15 \end{aligned}$$
>
> The E-value for the point estimate is 3 and for the lower confidence bound is 2.15. This means that an unmeasured confounder Z that is associated with **both** the outcome and exposure by a risk ratio of 3 could explain away the association, or one with risk ratios of at least 2.15 could make the results no longer statistically significant, but weaker confounding could not.

> Create a plot of the values of RR_{UA} and RR_{UY} that could explain away the observed association between the exposure and outcome. The following re-arranged version of equation 2 may be helpful:
>
> $$RR_{AU} = \frac{BF \times (1 - RR_{UY})}{BF - RR_{UY}}$$

```r
# returns RRau for given RRuy and BF
bounding_func <- function(RRuy, BF) {
  (BF * (1 - RRuy)) / (BF - RRuy)
}

RR <- 1.8
```

```r
LCI <- 1.4
Eval <- RR + sqrt(RR * (RR - 1))
EvalCI <- LCI + sqrt(LCI * (LCI - 1))

# propose values
RRuy <- seq(1, 40, by = .01)
# start off with empty vectors to be filled
RRau_RR <- rep(NA, length(RRuy))
RRau_LCI <- rep(NA, length(RRuy))
# run through all the RRuy values, calculating RRau for each
for (i in 1:length(RRuy)) {
  RRau_RR[i] <- bounding_func(RRuy[i], RR)
  RRau_LCI[i] <- bounding_func(RRuy[i], LCI)
}

# makde dataframes for each with identical colnames
bounds <-
  data.frame(RRau = RRau_RR, RRuy = RRuy, Bound = "Point Estimate")
CIbounds <-
  data.frame(RRau = RRau_LCI, RRuy = RRuy, Bound = "Lower Confidence Bound")

# put them together
plot_data <- rbind(bounds, CIbounds)

# we know the RRau has to be > 1
plot_data <- subset(plot_data, RRau > 1)

# make a dataframe to plot the points for the E-values
point_data <-
  data.frame(
    x = c(Eval, EvalCI),
    y = c(Eval, EvalCI),
    Bound = c("Point Estimate", "Lower Confidence Bound")
  )

# make labels for the points
label1 <- paste0("(", round(Eval, 2), ", ", round(Eval, 2), ")")
label2 <- paste0("(", round(EvalCI, 2), ", ", round(EvalCI, 2), ")")

ggplot(plot_data, aes(RRau, RRuy, group = Bound, col = Bound)) + geom_line() +
  geom_point(data = point_data, aes(x, y)) +
  annotate(
    "text",
    label = label1,
    x = Eval + 0.5,
    y = Eval + 0.3,
    size = 4
  ) +
  annotate(
    "text",
    label = label2,
    x = EvalCI + 0.5,
    y = EvalCI + 0.3,
```

```
  size = 4
) +
ylab(expression(RR[UY])) + xlab(expression(RR[AU])) + ylim(1, 10) + xlim(1, 10) +
scale_color_discrete(labels = c(
  expression(RR[AU] * RR[UY] * "/(" * RR[AU] * " + " * RR[UY] * " - 1) = 1.8"),
  expression(RR[AU] * RR[UY] * "/(" * RR[AU] *
               " + " * RR[UY] * " - 1) = 1.4")
)) +
theme(legend.position = "bottom")
```



4 Extensions

4.1 Protective associations

You may have noticed that it would be impossible to calculate an E-value in the case that the observed association is protective (as the term inside the square root in question 3 would be negative). However, we can easily accomodate protective associations by first transforming the ratio to the positive scale by taking the inverse.

For example, say that we conducted an observational study in which, after adjusting for multiple potential confounders, we found a protective association between exposure and outcome of $RR = 0.75$. If we were concerned about residual confounding we could calculate an E-value by the following procedure:

1. Take the inverse $RR^* = \frac{1}{RR} = \frac{1}{0.75} = 1.33$.

2. Plug this positive value in to equation 3:

$$\text{E-value} = RR^* + \sqrt{RR^*(RR^* - 1)} = 1.33 + \sqrt{1.33(1.33 - 1)} = 2$$

As before, this value represents the minimum strength of association on the risk ratio scale that an unmeasured confounder would need to have with both the exposure and the outcome, conditional on the measured covariates, to fully explain away a specific exposure-outcome relationship.

4.2 Prevalence specification

If we are willing to make a few additional assumptions and specify the prevalence of the unmeasured confounder U, then we can calculate the exact amount of bias (rather than just bounds). This could be useful if we wanted to calculate "corrected" effect measures.

Schlesselman showed that if there is no unmeasured confouding given (C, U), i.e. $Y_a \perp\!\!\!\perp A \mid (C, U)$ and U is single binary confounder with same risk ratio $\gamma = \frac{P(Y=1|U=1,A=a,C=c)}{P(Y=1|U=0,A=a,C=c)}$ for Y among exposed and unexposed then

$$B_{mult}(c) = \frac{1 + (\gamma - 1)P(U = 1 \mid A = 1, C = c)}{1 + (\gamma - 1)P(U = 1 \mid A = 0, C = c)}$$

Once we have calculated the bias factor $B_{mult}(c)$ we can invoke equation 1 and estimate a "corrected" risk ratio by dividing the observed estimate by $B_{mult}(c)$. Using similar logic we could also obtain corrected confidence intervals by dividing both limits by the bias factor $B_{mult}(c)$.

4.3 Other effect measures

In some cases another effect measure, other than the risk ratio, may be preferred. Vanderweele and Ding provide approximations that can be used for most other effect measures, to generate values for the RR that can then be plugged into equation 3 to calculate E-values.

Effect measure	Conditions	Approximation Point Estimate	Approximation Confidence Interval
OR	Prevalence of Y is less than 15%	$RR = OR$	same as point estimate
OR	Prevalence of Y is less than 15%	$RR = \sqrt{OR}$	same as point estimate
HR	Prevalence of Y is less than 15%	$RR = HR$	same as point estimate
HR	Prevalence of Y is less than 15%	$RR = \frac{1-0.5^{\sqrt{HR}}}{1-0.5^{1/\sqrt{HR}}}$	same as point estimate
IRR	In all cases	$RR = IRR$	same as point estimate
Cohen's d	In all cases	$RR = e^{0.91d}$	$(e^{0.91d-1.78s_d}, e^{0.91d+1.78s_d})$
RD	p_1 and p_0 between 0.2 and 0.8	$RR = e^{0.91RD}$	
RD	p_1 and p_0 not between 0.2 and 0.8	$RR = \frac{p_1}{p_0}$	complicated

5 Limitations

While the E-value is a simple and valuable tool for estimating how much residual confouding would explain away results, there are several important limitations that users should be aware of:

- E-values only provide information about residual confounding, they tell you nothing about the potential influences of other common forms of bias[1] like selection bias, measurement error, information bias, and

[1] E-value like quantities for some of these do exist now if they are a major concern

a number of other time biases. In the real world it is rare that a study suffers from residual confouding alone.

- E-values are not a substitute for deeper reflection about the unique sources of bias in any study. A large E-value means nothing if you've left out some very important known confounders (e.g. smoking) and likewise small E-values are only interpretable relative to the problem at hand.
- Don't forget that, for the sake of simplicity of presentation, the E-value assumes the case in which RR_{UA} and RR_{UY} are equal; however in practice these two quantities will rarely be equivalent. As you've learned this means that a counfounder with RR_{UY} that is smaller than the E-value could explain away the observed association but that it's RR_{UY} would have to be bigger than the E-value.
- While the E-value is still valid in circumstances in which multiple unobserved confounders are possible, it may be harder to conceptualize whether the amount of residual confouding implied by the E-value is likely or not because it now measures the "composite" effect of all unobserved confounders.

6 Resources for calculating bounds and E-values

You may find the following resources useful for your homework or for applying some of these concepts in your own research:

- There are packages in R (https://cran.r-project.org/web/packages/EValue/index.html) and Stata (https://ideas.repec.org/c/boc/bocode/s458592.html) that contain some handy functions for performing E-value and sensitivity analysis calculations.
- There is an online calculator at https://www.evalue-calculator.com where you can calculate E-values and corresponding graphs for a number of observed effect measures.

7 Appendix

7.1 Derivation of E-value

$$\begin{aligned} RR_{obs} &= BF \cdot RR_{true} \\ RR_{obs} &= \frac{RR_{eq}^2}{2RR_{eq} - 1} \cdot (1) \\ RR_{obs}(2RR_{eq} - 1) &= RR_{eq}^2 \\ 2RR_{obs}RR_{eq} - RR_{obs} &= RR_{eq}^2 \\ 0 &= RR_{eq}^2 - 2RR_{obs}RR_{eq} + RR_{obs} \end{aligned}$$

Using quadratic equation

$$\begin{aligned} RR_{eq} &= \frac{2RR_{obs} \pm \sqrt{4RR_{obs}^2 - 4RR_{obs}}}{2} \\ &= \frac{2RR_{obs} \pm 2\sqrt{RR_{obs}^2 - RR_{obs}}}{2} \\ &= RR_{obs} \pm \sqrt{RR_{obs}(RR_{obs} - 1)} \end{aligned}$$

The positive root $RR_{obs} + \sqrt{RR_{obs}(RR_{obs} - 1)}$ is the only one that makes sense here so we define it as the E-value