Chapter 3

Observational Studies

3.1 Causal Inference

Causal inference means using results from a statistical analysis to make inference about the causal effect of a treatment/intervention/exposure on an outcome.

Notation

- Let Y be the outcome
- Let Z be the treatment or exposure, where Z is binary e.g., $Z = 1$ denotes treatment and $Z = 0$ denotes control
- Let X be some baseline or pre-treatment covariates
- We will use potential outcomes notation:
  - $Y^{(1)}$ is the outcome when $Z = 1$ e.g., under treatment
  - $Y^{(0)}$ is the outcome when $Z = 0$ e.g., under control

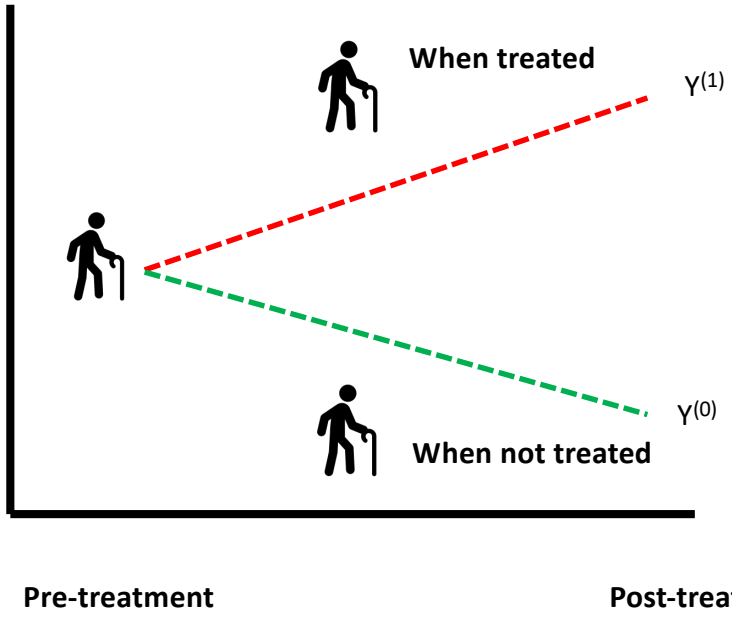
Everyone in the study has a potential $Y^{(1)}$ and $Y^{(0)}$, regardless of what treatment they actually received.

However, only one of $Y^{(1)}$ or $Y^{(0)}$ is actually observed.

The **causal effect** is the difference between these two potential outcomes i.e., $Y^{(1)}$ - $Y^{(0)}$, illustrated in Figure 3.1.

Primary types of causal effects

Figure 3.1: Causal effect of a treatment



- The **Average Treatment Effect in the Population** (ATE) answers the question: How effective is the treatment in the population?
  - ATE $= E(Y^{(1)} - Y^{(0)})$
- The **Average Treatment Effect in the Treated** (ATT) answers the question: How would those who received treatment have done had they received the comparison condition?
  - ATT $= E(Y^{(1)} - Y^{(0)}|Z = 1)$
- The ATE is of more interest if every treatment potentially might be offered to every subject, whereas the ATT is preferable when a patient's characteristics are more likely to determine the treatment received.
- We will focus on the ATE.

Randomized Studies

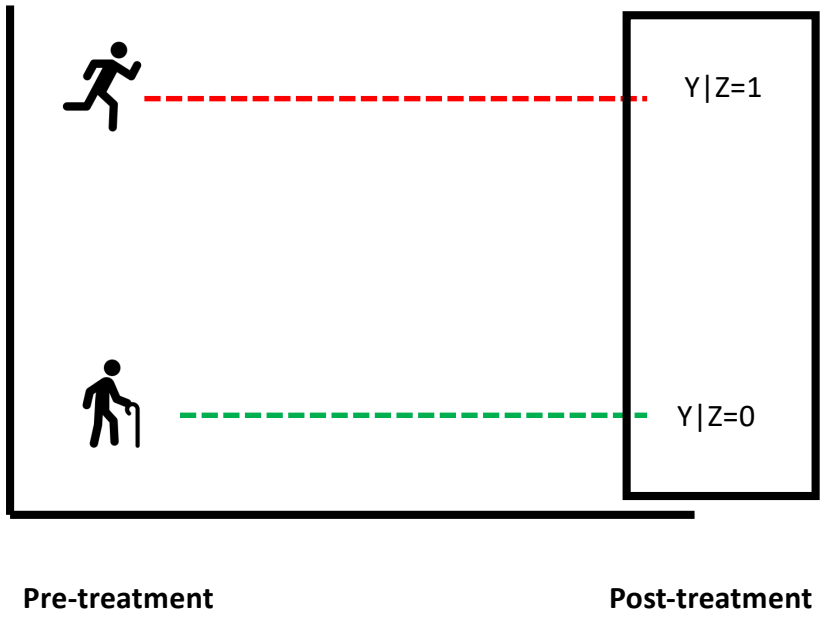
- Random experiments = gold standard for estimating causal effects
- Randomization (if it works) makes the two groups that are being compared **balanced** on baseline characteristics

- $\Rightarrow$ Treatment assignment is unrelated to potential outcomes (strong ignorability)
- Randomization is not always feasible

Observational Studies

- Observational studies provide another way to get at causal effects
- Treatment assignment is not controlled by the researcher
- Groups being compared are usually **imbalanced**
- Can use causal inference methods to try to replicate what a randomized study does

Figure 3.2: Selection bias

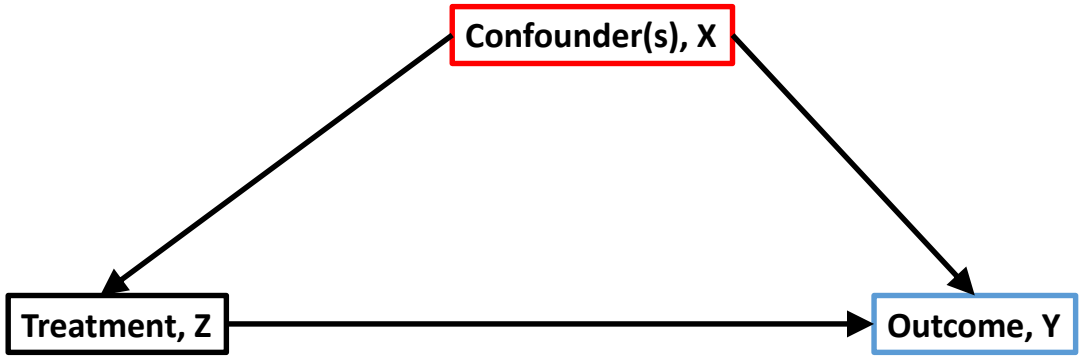


A significant challenge in causal effect estimation is **selection bias.**

- Selection occurs when the individuals who receive treatment differ from the individuals who receive control to begin with.
- Selection bias occurs when these differences are related to the outcome being measured (see Figure 3.2).

- A characteristic that is both associated with treatment "selection" and associated with the outcome is called a **confounder.** The existence of a confounder that is not appropriately accounted for in the analysis can result in selection bias.
- In causal inference, a directed acyclic graph (DAG) is often used to visualize relationships between variables. Figure 3.3 shows a simple DAG reflecting the existence of a confounder.

Figure 3.3: Directed Acyclic Graph



3.2 Propensity Scores

There are multiple approaches to handle selection bias. One popular approach is simply **regression adjustment** i.e., adjusting a regression examining the effect of Z on Y for X. This is valid as long as the regression model is correctly specified and X includes all confounders.

Another approach is via the use of propensity scores. The **propensity score** is an individual's probability of receiving treatment given pre-treatment characteristics, denoted as $p(X)$:

$$p(X) = P(Z = 1|X)$$

Propensity scores are used to create **balance** between the treatment and control group. Propensity scores can be used to estimate causal effects if the following hold:

1. $Y^{(1)}, Y^{(0)} \perp Z|X \Rightarrow$ there are no unmeasured confounders

2. $0 < p(X) < 1$, $\Rightarrow$ there is overlap between groups

Once you have the estimated propensity scores there are different ways to use them to estimate causal effects:

- Stratification
- Matching
- Weighting $\Rightarrow$ in this class, **we will focus on propensity score weighting**

Specifically, it can be shown that the ATE can be estimated using propensity score weighting as:

$$\widehat{ATE} = \frac{\sum_{i:Z_i=1} Y_i W_i}{\sum_{i:Z_i=1} W_i} - \frac{\sum_{j:Z_i=0} Y_j W_j}{\sum_{j:Z_i=0} W_j}$$

where

$$W_i = \begin{cases} 1/p(X_i) & \text{if} \quad Z_i = 1 \\ 1/(1-p(X_i)) & \text{if} \quad Z_i = 0 \end{cases}$$

This quantity weights each person depending on their likelihood of receiving treatment. Note that the incorrect estimate of the ATE would be

$$\widetilde{ATE} = \frac{1}{n_1} \sum_{i:Z_i=1} Y_i - \frac{1}{n_0} \sum_{j:Z_i=0} Y_j$$

where n_1 is the number of people in the treatment group ($Z_i = 1$) and n_0 is the number of people in the control group ($Z_j = 0$).

Propensity Score Estimation and Balance

How do we estimate the propensity score? The standard approach has been to use a logistic regression model with the treatment indicator as the outcome and the confounders as predictors. However, this is not ideal due to potential model misspecifiation.

Machine learning methods have been shown to perform better than logistic regression in propensity score estimation

- Achieve better balance between treatment and comparison group on pretreatment covariates
- Reduce bias in treatment effect estimates
- Produce more stable propensity score weights (thereby also improving precision)

There are many methods to estimate propensity scores, some examples are:

- Covariate Balancing Propensity Score (Imai and Ratkovic 2013)
- Super Learning (van der Laan 2014)
- High Dimensional Propensity Score (HDps) (Schneeweiss et al 2009)
- Entropy Balance (Hainmueller 2012)

We will focus on using **generalized boosted models (GBM)** to estimate propensity score weights.

GBM is a **nonparametric approach** to model outcomes (binary, discrete, or continuous) that allows for interactions among covariates and flexible functional forms for the regression surface. It is also **invariant to monotonic transformations** of covariates.

- GBM models the log odds of treatment assignment

$$g(X) = \log\left\{\frac{p(X)}{1-p(X)}\right\}$$

  by initially setting $g(X)$ to $\log(\bar{z}/(1-\bar{z}))$, the constant baseline log odds of assignment to the treatment, where $\bar{z}$ is the average treatment assignment indicator for the entire sample.

- The next step of the algorithm searches for a small adjustment, $h(X)$, to add to the initial estimate and improve the fit of the model to the data.

- Fit is measured by the Bernoulli log-likelihood:

$$l(g) = \sum_i z_i g(X_i) - \log(1 + \exp(g(X_i)))$$

  with larger values implying better fit, where $z_i = 0$ when $g(X_i)$ is negative and $z_i = 1$ when $g(X_i)$ is positive.

- If the algorithm finds an adjustment that improves the fit, then the current model becomes $g(X) + h(X)$. The boosting procedure iterates, each time selecting a model adjustment that when added to $g(X)$ offers an increase in the log- likelihood.

- Technically, $h(X)$ can be of any form, but here, $h(X)$ is selected to be a regression tree that models the residuals from the current fit as a function of the covariates.

Heuristically, GBM models an outcome as a **sum of simple regression tree fits** and each iteration of the fitting algorithm adds an additional tree fit to the residuals of the model from the previous iteration.

Given the sum-of-trees formulation, the number of iterations used in the fitting algorithm is commonly referred to as the **"number of trees"** in the model. The GBM algorithm

improves the fit to the data with each additional tree, requiring an external criterion to select the number of trees that is optimal in a given situation and controls between overfitting to the data and underspecification of the model.

For propensity score estimation, the **balance of the covariates across treatment and control groups, that is, the similarity in the weighted distributions of covariates from the two groups, is used to select the optimal number of trees in the GBM.**

What do we mean by balance? Suppose you have a treatment and control group and the mean age of individuals in the treatment group is 22 while the mean age in the control group is 65. These groups are **not balanced.**

Suppose you have a treatment and control group and 80% of the treatment group is female while 30% of the control group is female. These groups are **not balanced.**

Suppose you have a treatment and control group and 80% of the treatment group is female while 85% of the control group is female. Are the groups balanced?

Metrics to Assess Balance

Commonly used metrics to assess balance:

- p-value for a t-test (or equivalent) between the two groups
- standardized effect size (ES) difference between the two groups:
  - the difference in the mean of a pretreatment variable in the treated group versus the control group, divided by the standard deviation
  - threshold of 0.2 is often used
- Kolmogorov-Smirnov (KS) statistic:
  - the maximum vertical distance between the empirical cumulative distribution functions of two samples

We will use the `twang` package in `R` to estimate propensity score weights, assess balance, and estimate the ATE. The `twang` package which stands for Toolkit for Weighting and Analysis of Nonequivalent Groups, uses GBM for propensity score estimation and selects the optimal number of trees as that which maximizes balance (as measured by either ES or KS) between the treatment groups being considered.

The main workhorse of `twang` is the `ps()` function which implements generalized boosted regression modeling to estimate the propensity scores. This software is available in R, SAS and STATA.

Steps to estimate the treatment effect

1. Identify potential confounders
2. Assess balance without propensity score weighting
3. Estimate propensity scores
4. Assess balance of weighted groups
5. If balance is achieved, estimate the treatment effect using the propensity score weighted estimator
6. Interpret within the context of required assumptions

References:

Ridgeway, G., McCaffrey, D. F., Morral, A. R., Cefalu, M., Burgette, L. F., Pane, J. D., & Griffin, B. A. (2022). Toolkit for weighting and analysis of nonequivalent groups: a tutorial for the R TWANG package. Santa Monica, Calif: Rand. `https://cran.microsoft.com/snapshot/2017-07-10/web/packages/twang/vignettes/twang.pdf`

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in medicine, 34(28), 3661-3679.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3), 399-424.

3.3 Double Robust Estimation

To estimate a causal effect, we have discussed two options (1) regression adjustment and (2) propensity score weighting.

Both of these rely on the assumption that there are **no unmeasured confounders** $\Rightarrow Y^{(1)}, Y^{(0)} \perp Z|X$

Regression adjustment will result in a valid treatment effect estimate if the regression model or, **the outcome model** is correctly specified. If the model is not correct, you may get a biased estimate.

- Suppose you postulate the following model and that this model is correct:

$$E(Y) = \beta_0 + \beta_1 Z + \beta_2 X$$

  That is, you fit the linear regression model: $Y \sim Z + X$.

- Under the assumption of no unmeasured confounders, it can be shown that:

$$\text{ATE} = E(Y^{(1)} - Y^{(0)}) = \beta_1$$

  Thus, when you fit the model $Y \sim Z + X$, you obtain an estimate of β_1 and this is a valid estimate of the ATE.

- However, let's say that in fact the true model is

$$E(Y) = \alpha_0 + \alpha_1 Z + \alpha_2 X^2$$

  By similar algebra, you can show that ATE $= \alpha_1$. In general $\beta_1 \neq \alpha_1$. So if you fit the model $Y \sim Z + X$, you will get an estimate of β_1, not α_1. Therefore, your estimate of ATE will be incorrect.

Propensity score weighting will result in a valid treatment effect estimate if **the propensity score model**, the model you used to estimate the propensity scores, is correctly specified. If the model is not correct, you may get a biased estimate.

- Recall our propensity score weighted estimate of ATE:

$$\widehat{ATE} = \frac{\sum_{i:Z_i=1} Y_i W_i}{\sum_{i:Z_i=1} W_i} - \frac{\sum_{j:Z_i=0} Y_j W_j}{\sum_{j:Z_i=0} W_j}$$

  where

$$W_i = \begin{cases} 1/p(X_i) & \text{if} \quad Z_i = 1 \\ 1/(1-p(X_i)) & \text{if} \quad Z_i = 0 \end{cases}$$

- This is called the **inverse probability weighted (IPW)** estimate of the treatment effect.
- If the model used to obtain the propensity scores: $p(X) = P(Z = 1|X)$ is correct, then under the assumption of no unmeasured confounders, it can be shown that:

$$E(\widehat{ATE}) = E(Y^{(1)} - Y^{(0)}) = ATE$$

- However, if the model for $p(X)$ is not correct e.g., $P(Z = 1|X) = p^*(X) \neq p(X)$, then $E(\widehat{ATE}) \neq ATE$ and your estimate will be incorrect.

Double robust estimation is an estimation method that essentially combines these two ideas resulting in a valid estimate as long as **either** the propensity score model or the outcome model is correctly specified; see Figure 3.4.

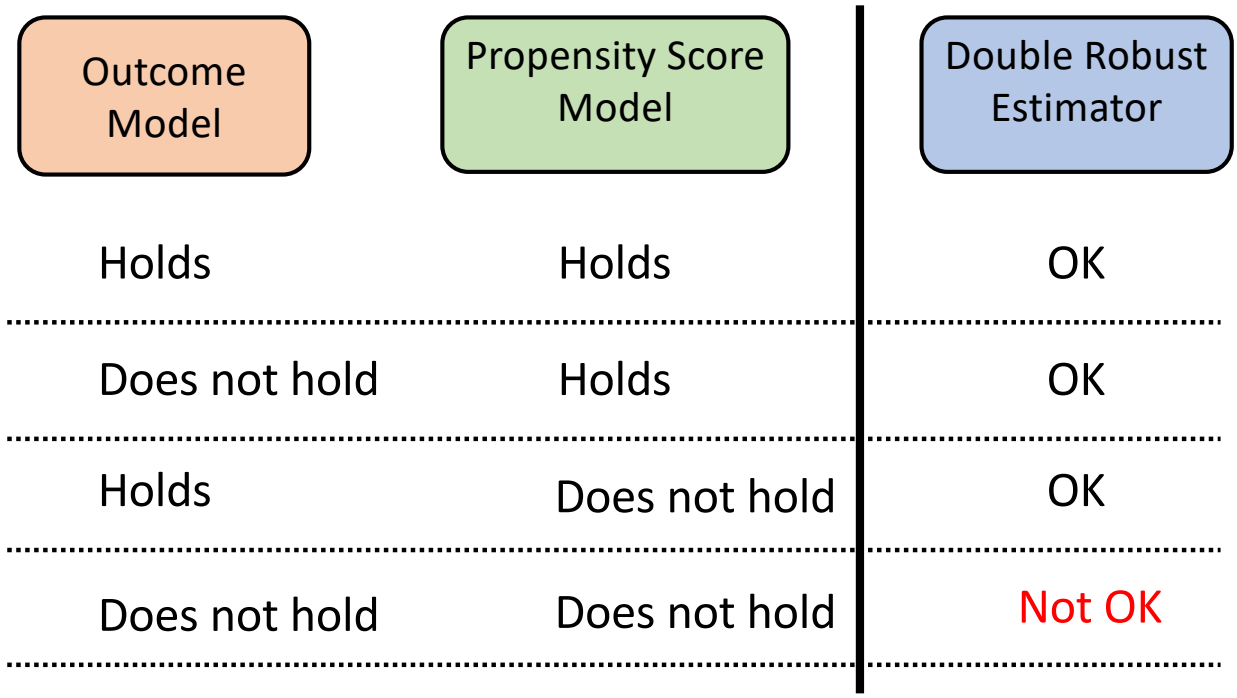


Figure 3.4: Double Robust Estimator

The concept of "double robustness" is not unique to treatment effect estimation; the estimator we will discuss here is one example of a double robust estimator. In general, it is an estimate or an approach that will be valid as long as A or B hold, where A and B can be models or assumptions. You don't have to know which one is correct.

There are different ways to combine the outcome model and the propensity score model to create a double robust estimator.

One example is:

$$\begin{aligned}\widehat{ATE}_{DR} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i Y_i}{p(X_i)} - \frac{\{Z_i - p(X_i)\}}{p(X_i)} m(X_i, Z_i)\right] \\ &- \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-Z_i)Y_i}{1-p(X_i)} - \frac{\{Z_i - p(X_i)\}}{1-p(X_i)} m(X_i, Z_i)\right] \\ &= \widehat{\mu}_{1,DR} - \widehat{\mu}_{0,DR}\end{aligned}$$

where $p(X_i)$ is the estimate from the propensity score model and $m(X_i, Z_i)$ is the estimate of Y from the outcome model. For example, $p(X_i)$ may be obtained by simple logistic regression $Z \sim X$ and $m(X_i, Z_i)$ may be obtained using simple linear regression $Y \sim X + Z$.

- Each term, $\widehat{\mu}_{1,DR}$ and $\widehat{\mu}_{0,DR}$ can be viewed as taking an IPW estimate and "augmenting" it with a second component.
- It can be shown that if <u>either</u> $p(X)$ or $m(X, Z)$ is correct, then

$$E(\widehat{ATE}_{DR}) = E(Y^{(1)} - Y^{(0)}) = ATE$$

- If both models are wrong, then $E(\widehat{ATE}_{DR}) \neq ATE$.
- This double robust estimator:
  - Offers protection against model misspecification
  - Will have **smaller variance** (in large samples) than the simple IPW estimator when the model for $p(X)$ is correct
  - Will have **larger variance** (in large samples) than the simple outcome model estimator when the model for $m(X, Z)$ is correct - but gives protection in the event that it is not correct.

How do you decide which X to include in the models? It is not the case that throwing everything you have in the models is the right thing to do. For IPW estimators:

- Variables unrelated to exposure but related to outcome should always be included in the propensity score model $\Rightarrow$ increased precision
- Variables related to exposure but unrelated to outcome can be omitted $\Rightarrow$ decreased precision

References:

Davidian, M. Double Robustness in Estimation of Causal Treatment Effects. `https://www4.stat.ncsu.edu/~davidian/double.pdf`

Lunceford, J.K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Statistics in Medicine 23:2937–2960

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics 61, 962–972.

Kang, JDY, and Schafer, JL. (2007) Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science. 22(4): 523-539.

Brookhart, M. A. et al. (2006). Variable selection for propensity score models. American Journal of Epidemiology 163, 1149–1156.

Brookhart, M. A. and van der Laan, M. J. (2006). A semiparametric model selection criterion with applications to the marginal structural model. Computational Statistics and Data Analysis 50, 475–498.

See Tan, Z. (2006) A distributional approach for causal inference using propensity scores. Journal of the American Statistical Association 101, 1619–1637.

3.4 Unmeasured Confounding

The assumption of unmeasured confounding never holds in an observational study. There is always going to be something that affects both the treatment and the outcome that is not measured or not measurable.

The question is not whether it holds. The question is - does there exist an unmeasured confounder that, if you measured it and accounted for it, would change your conclusion about the treatment effect?

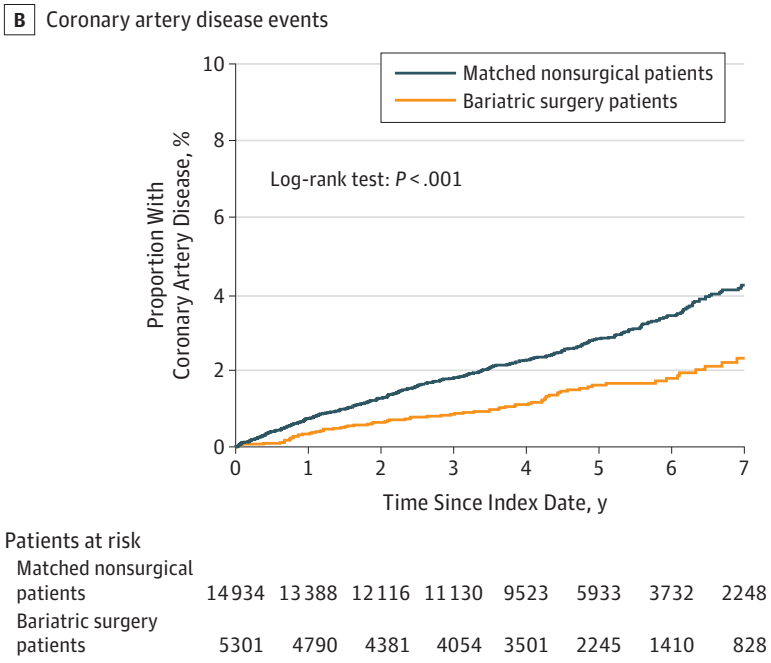


Figure 3.5: Cumulative Incidence Rates from JAMA 2018 Publication

Consider an example from a JAMA 2018 paper:

- In the October 16, 2018, issue of JAMA, results from a large, multisite observational study of the association between bariatric surgery and long-term macrovascular disease outcomes among patients with severe obesity and type 2 diabetes was reported.
- Using data from **5,301 patients** aged 19 to 79 years who underwent bariatric surgery at 1 of 4 integrated health systems in the United States between 2005 and 2011 and **14,934** matched nonsurgical patients, they found that bariatric surgery was associated with lower incidence of macrovascular disease at 5 years (HR = 0.60, 95% CI: 0.42, 0.86).
- Two strategies were used to mitigate confounding bias. In the first, **a matched cohort design** was used where nonsurgical patients were matched to surgical patients on

the basis of a priori–identified potential confounders (study site, age, sex, body mass index, hemoglobin A1c level, insulin use, observed diabetes duration, and prior health care use). In the second strategy used to adjust for confounding bias, the primary results were based on the fit of a **multivariable Cox model** that adjusted for all of the factors used in the matching as well as a broader range of potential confounders (regression adjustment). Thus, any imbalances in the observed potential confounders that remained after the matching process were controlled for by the statistical analysis.

- Despite these efforts, however, given the observational design, the potential for unmeasured confounding remained. You cannot adjust for things that are not measured.

The E-value

An E-value analysis asks the question: **how strong would the unmeasured confounding have to be to negate the observed results?**

- The E-value itself answers this question by quantifying the **minimum strength of association on the risk ratio scale that an unmeasured confounder must have with both the treatment and outcome**, while simultaneously considering the measured covariates, to negate the observed treatment-outcome association.
- E-values can help assess the robustness of the main study result by considering whether unmeasured confounding of this magnitude is plausible.
- The E-value provides a measure related to the evidence for causality, hence the name "E-value."

The E-value has some appealing features:

1. It is intuitive because the lowest possible number is 1. The higher the E-value is, the stronger the unmeasured confounding must be to explain the observed association.
2. The E-value is simple to calculate for a range of effect measures, including relative risks, HRs, and risk differences, and study designs. The formulas for the E-value for different effect measures, including continuous outcomes, are available and the E-value has been implemented in freely available software and an online calculator
3. The calculation can also be applied to the bounds of a 95% CI. That is, you can assess the extent of unmeasured confounding that would be required to shift the confidence interval.

What is the E-value?

Let's take an example with a binary outcome, denoted as D, and binary exposure, denoted as E. Suppose you have a set of measured confounders, denoted as C, and that there exists a set of unmeasured confounders, denoted as U.

Let RR_{ED}^{true} denote the true relative risk of the exposure E on the outcome D, adjusted for **both** C and U, measured and unmeasured confounders.

In contrast, let RR_{ED}^{obs} denote the observed (estimated) relative risk of the exposure E on the outcome D, adjusted for **measured confounders**, C and U.

Let RR_{EU} be the relative risk of the exposure E on the unmeasured confounder (technically, the maximum RR among U), and let RR_{UD} be the relative risk of the unmeasured U on the outcome D (technically, the maximum RR among U and stratified by exposure).

It can be shown that:

$$RR_{ED}^{true} \geq RR_{ED}^{obs} / \frac{RR_{EU} RR_{UD}}{RR_{EU} + RR_{UD} - 1}.$$

This means that even in the presence of unmeasured confounding, the true relative risk must be at least as large as

$$RR_{ED}^{obs} / \frac{RR_{EU} RR_{UD}}{RR_{EU} + RR_{UD} - 1}$$

and the quantity $\frac{RR_{EU} RR_{UD}}{RR_{EU} + RR_{UD} - 1}$ is a "joint bounding factor" for the relative risk.

The quantities RR_{EU} and RR_{UD} are considered sensitivity parameters. If these two parameters are taken to be equal, the E-value is the minimum value for both associations that would be capable of attenuating the observed association to the null.

The **E-value can be calculated** for an observed risk ratio, RR_{ED}^{obs}, can be obtained as

$$\text{E-value} = RR_{ED}^{obs} + \sqrt{RR_{ED}^{obs}(RR_{ED}^{obs} - 1)}.$$

If the observed risk ratio is below 1, then one first takes the inverse before applying the E-value formula.

This formula can also be used for hazard ratios or odds ratios with outcomes that are rare at the end of follow-up. For hazards or odds ratio with a common outcome at the end of follow-up, or with continuous outcomes, approximate E-values can still be obtained through various transformations

How to calculate the E-value?

R package: `EValue`
Website: `https://www.evalue-calculator.com`

Interpreting the E-value

The E-value for the example above was 2.72, meaning that residual confounding could explain the observed association if there exists an unmeasured covariate having a relative risk association at least as large as 2.72 with both macrovascular events and with bariatric surgery.

The HRs for some of the known, powerful macrovascular disease risk factors were 1.09 for hypertension, 1.88 for dyslipidemia, and 1.48 for being a current smoker. It is not likely that an unmeasured or unknown confounder would have a substantially greater effect on macrovascular disease development than these known risk factors by having a relative risk exceeding 2.72.

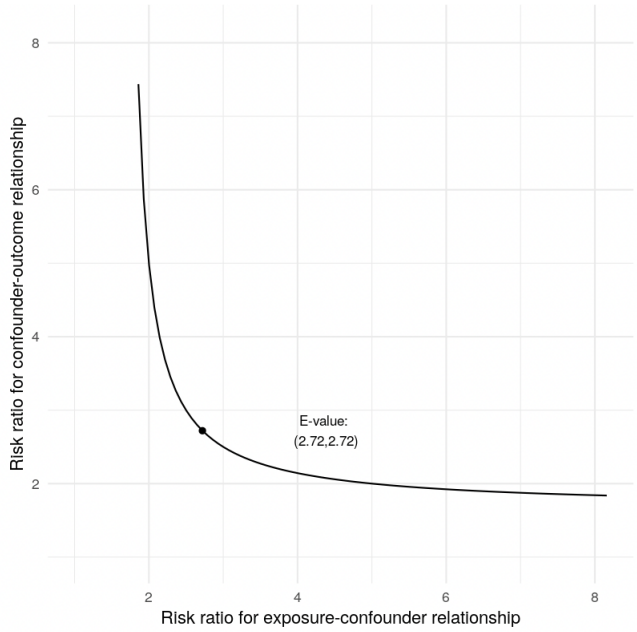


Figure 3.6: E-value plot using JAMA 2018 article results

Additional Example

Consider this study: Muhsen et al. (2022). Association of receipt of the fourth BNT162b2 dose with omicron infection and COVID-19 hospitalizations among residents of long-term care facilities. JAMA internal medicine, 182(8), 859-867. `https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/2793699`

The purpose of this study was to determine the association of the fourth BNT162b2 dose with protection against SARS-CoV-2–related infections, hospitalizations, and deaths during the Omicron surge in long-term care facility (LTCF) residents.

Select one of the outcomes: **What is the E-value? What is the conclusion?**

References:

Haneuse, S., VanderWeele, T. J., & Arterburn, D. (2019). Using the E-value to assess the potential effect of unmeasured confounding in observational studies. Jama, 321(6), 602-603.

Ding, P., & VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. Epidemiology (Cambridge, Mass.), 27(3), 368.

VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. Annals of internal medicine, 167(4), 268-274.