Confounding and beyond: Non-collapsibility, colliders

Propensity Score Methods

Lecture 7 and Lecture 8

Background

- In medical research it is very common to focus on the estimation of the effect of intervention, policies or treatments.
- In randomized experiments, the randomization enables unbiased estimation of treatment effects
  - Randomization implies that treatment groups will be balanced on average with respect to both observed and unobserved risk factors

Background

- There are cases, however, where researchers focus on estimating the effect of interventions with respect to which it is not possible or ethical to randomize.
- Example: A researcher may want to estimate the effect of smoking.

Background

- The possibility of bias arises because a difference in the treatment outcome (such as the average treatment effect) between treated and untreated groups may be caused by a factor that predicts treatment rather than the treatment itself (confounding).
- Statistical methods aim on removing the confounding effect
- In the following we will focus on available statistical methods and when to use each of them

The Simpson's Paradox (1951)

Numerical example: Three dichotomous variables, let's say corresponding to exposure (A), disease (B) and an individual characteristic (C) measured in a population of N=52 individuals.

	Exposed	Unexposed
Disease	20	20
No disease	6	6

$$OR = \frac{\frac{20}{6}}{\frac{20}{6}} = 1$$

	C=1		C=0	
	Exposed	Unexposed	Exposed	Unexposed
Disease	5	8	15	12
No disease	3	4	3	2

$$OR = \frac{\frac{5}{8}}{\frac{3}{4}} = \frac{5}{6}$$

$$OR = \frac{\frac{15}{12}}{\frac{3}{2}} = \frac{5}{6}$$

To adjust or not adjust?

Suppose that A represents some medical treatment (1: yes, 0: no), B represents death (1: yes, 0: no) and C represents individual's gender (1: male, 0: female).

In this setting, the investigator would be interested in the conditional odds ratio $OR_{AB|C}$, which shows that the treatment is associated with a lower risk of death in both men and women.

Hernán M, Clayton D, Keiding N. The Simpson's paradox unraveled. Int J Epidemiol. (2011)

Let's recall what confounding is...

Men were less likely to receive treatment (OR_{AC}=0.34), thus treatment A was not randomly assigned in the study population.

Men were less likely to die (OR_{BC}=0.52).

Treatment and death are associated conditional on gender.

A common cause like C **will create an association** between its effects A and B if not taken under consideration in the analysis (e.g by adjustment).

This association does not reflect the causal effect of A on B and is commonly referred to as confounding.

Counfounding=Common Cause

The causal diagram depicts the variables treatment (A), death (B) and sex (C).

A confounder C

Conditioning on variable C removes the confounding!

- In Simpson's example, the confounding results in a positive (OR>1) association between A and B because men are less likely to be treated and also less likely to die.
- There are two sources of association between treatment A and death B:
  - The positive association due to confounding by C and
  - The negative association presumably due to the protective effect of treatment A on the risk of death B.

- The unadjusted odds ratio (OR_{AB}=1) measures the combination of these two associations.

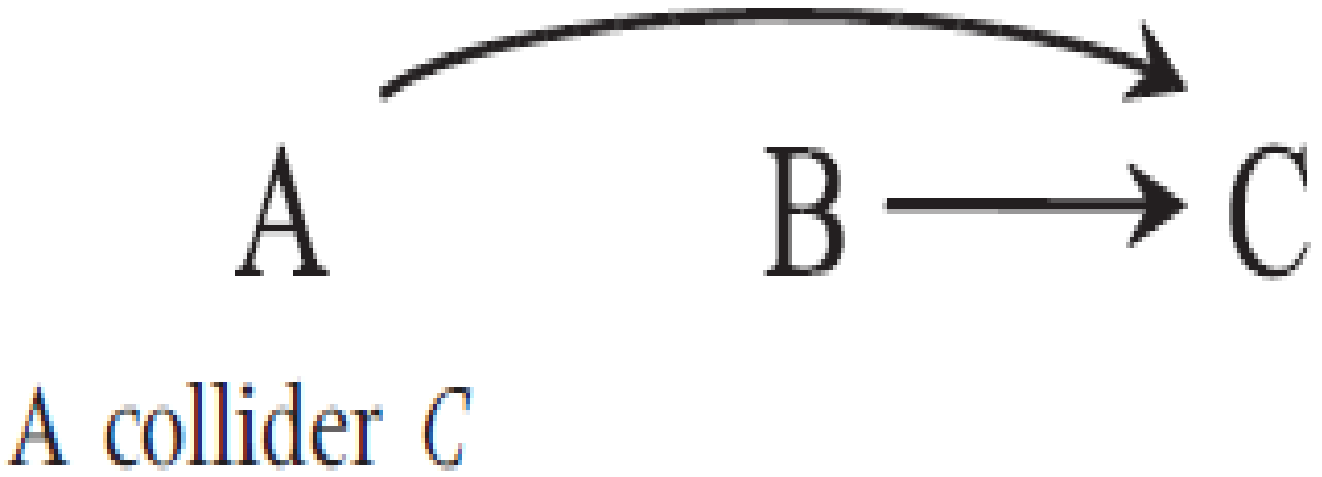
Confounding effect

- These are example data, generated in a way that the two effects cancel out and result in unadjusted OR=1
- In general, one would expect that the unadjusted odds ratio would be different from 1:
  - OR_{AB}>1 if the association due to confounding is greater than the association due to the effect of treatment on death
  - OR_{AB}<1 otherwise.

Same numbers, different setting

- Now, suppose that an investigator wished to examine whether in a pack of 52 cards the proportion of court cards (King, Queen, Knave) was associated with colour (Red, Black).
- However, a baby had been playing with the cards earlier and some of the cards were dirty.
- In this setting, A is the type of card (1: plain, 0: court), B the card's colour (1: black, 0: red) and C whether the card was dirty (1: yes, 0: no).
- In this setting, the investigator would be interested in the marginal odds ratio OR_{AB}, which is obviously 1 as a cards deck contains the same number of black and red court cards.

Collider=Common effect



A collider *C*

The common effect C is referred to as a collider because two arrowheads collide into it.

Collider

There is no arrow between A and B because card type and colour do not cause each other.

There is an arrow from A to C because, according to the example above, the baby had a strong predilection for court cards (OR_{AC}=0.34).

Similarly, there is an arrow from B to C because the baby preferred the red cards (OR_{BC}=0.52).

Conditioning on a collider C introduces bias!

- **Conditioning on a collider C generally introduces an association between its causes A and B even if the causes are marginally independent.**
- Therefore, the association that appears between A and B only when the analysis is conditional on C is expected.
  - Informally, if we select a court card that is known to be dirty, then it is less likely that the card is red.
- The baby appears to get cards dirty in such a way that the odds ratio $OR_{AB|C=1}$ among the dirty was exactly equal to the odds ratio $OR_{AB|C=0}$ among the clean, this is because the data are created this way.

Conditioning on a collider C introduces bias!

The association created between two variables by conditioning on their common effect is often referred to as selection bias.

Selecting a stratum of their common effect would have the same consequences.

Epidemiological example:

When estimating the effect of genetic factor A on diabetes B, one would generally introduce selection bias by conditioning on the history of heart disease C, since heart disease is an effect of both diabetes and the genetic factor

Which one would we report? Adjusted or Unadjusted?

Even if the conditional A–B association measure is homogeneous within levels of C, the sensible answer is sometimes the conditional association measure and other times the marginal one.

From a purely statistical standpoint, no general rule seems to exist as to whether the conditional association or the marginal association should be preferred.

Non-Collapsibility

We say a measure of association of X and Y is strictly collapsible across Z if it is constant across the strata (subtables) and this constant value equals the value obtained from the marginal table.

Consider a GLM for outcome Y with covarates W,X,Z:

$$g[E(Y \vee W, X, Z)] = \alpha + \beta w + \gamma x + \delta z$$

Omiting Z, the model becomes: $g[E(Y \vee W, X)] = \alpha' + \beta' w + \gamma' x$

If omiting Z $\beta = \beta^*$, then the regression is collapsible for β over Z

If omiting Z $\beta \neq \beta^*$, then the regression is not collapsible for β over Z

In the absence of confounding adjusted and unadjusted estimates of exposure effect in linear models will coincide. However, for non-linear models, even in the absence of confounding adjusted and unadjusted estimates will not coincide. This phenomenon is called non-collapsibilty and it is often confused with confounding

Confounding vs. Noncollapsibility

- Confounding: lack of exchangeability between exposed and non-exposed individuals due to inherent differences in their risk profiles [Greenland and Robins, 1986].
- Bias from confounding can be in any direction, and can lead to an observational association which is in the opposite direction to the true underlying causal association [Davey Smith and Ebrahim, 2002].
- In contrast, when there are no confounders, non-collapsibility may still alter the magnitude of an association, although it will not change its direction [Gail et al., 1984].

Estimating and contextualizing the attenuation of odds ratios due to non-collapsibility. Burges S. Communication in Statistics- Theory and Methods (2015)

Schuster, N.A., Twisk, J.W.R., ter Riet, G. et al. Noncollapsibility and its role in quantifying confounding bias in logistic regression. BMC Med Res Methodol 21, 136 (2021).

Exposure: not overweight vs. overweight

Outcome: diabetes

Covariate: sex

N= 200 individuals, 50% are overweight

80 individuals have diabetes and 120 individuals do not have diabetes.

Non-Collapsibility- Example

```
. logistic event exposure, or

Logistic regression                                     Number of obs     =        200
                                                        LR chi2(1)        =      19.11

Prob > chi2       =     0.0000
Log likelihood =  -125.0474                             Pseudo R2         =     0.0710

------------------------------------------------------------------------------
       event | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exposure |   3.666666   1.122607     4.24   0.000     2.012161    6.681591
       _cons |   .3333334   .0769801    -4.76   0.000     .2119836    .5241498
------------------------------------------------------------------------------
```

Non-Collapsibility- Example

Gender evenly distributed over the weight groups: both groups consist of 50 males and 50 females.

Weight status is not influenced by sex.

Gender and diabetes are associated: Of the 100 women, 30 have diabetes and 70 do not, whereas for males half have diabetes and half do not.

Non-Collapsibility- Example

Because confounding requires the covariate to be associated with both the exposure and the outcome, sex is not a confounder in the relation between weight and diabetes, as it is not associated with weight.

Since sex is not a confounder of the exposure-outcome effect, adjustment for sex should not affect the exposure-outcome effect estimate.

Non-Collapsibility- Example

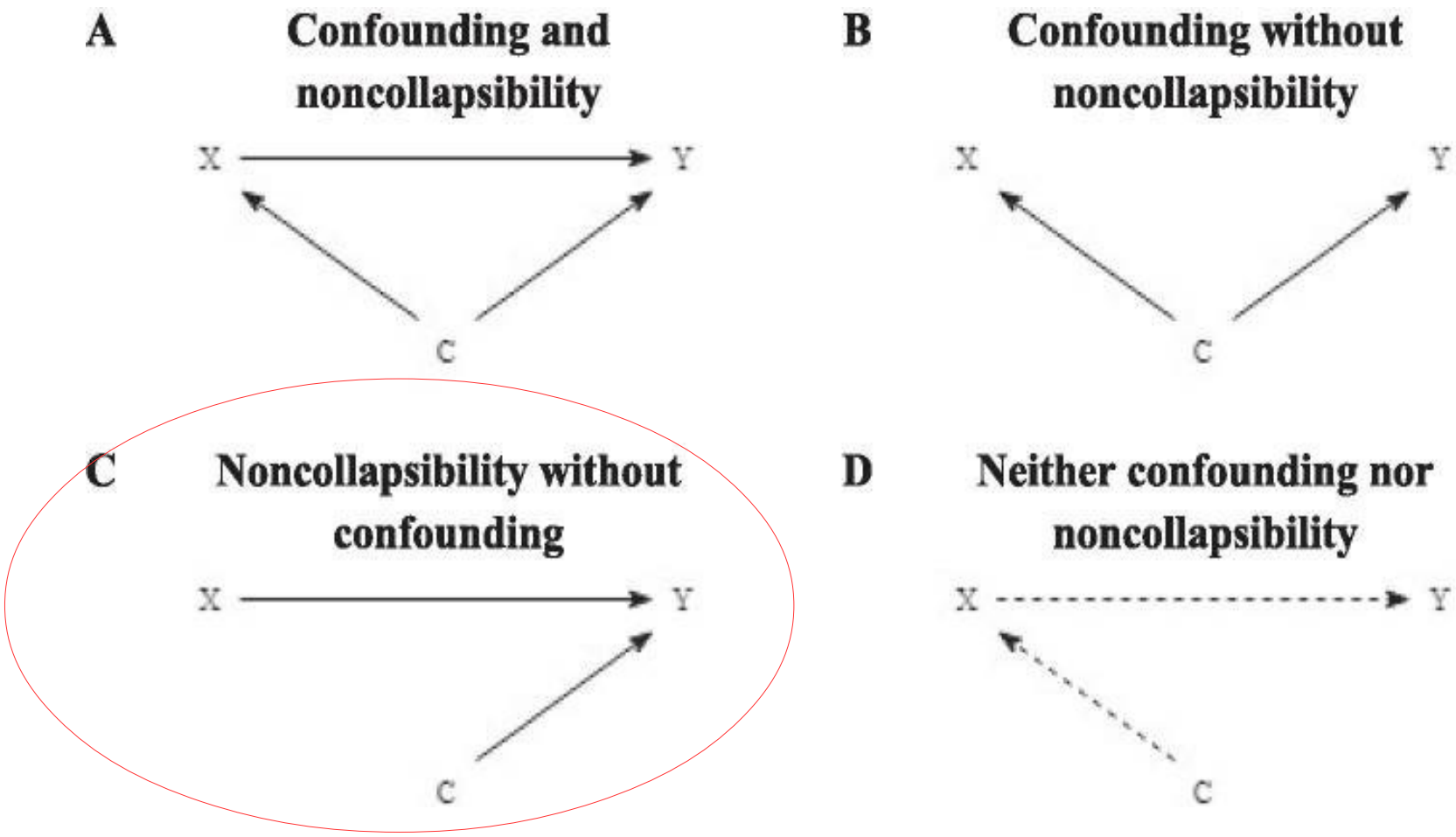
```
. logistic event exposure gender, or

Logistic regression                               Number of obs     =        200
                                                  LR chi2(2)        =      28.42
                                                  Prob > chi2       =     0.0000
Log likelihood = -120.39357                       Pseudo R2         =     0.1056

------------------------------------------------------------------------------
       event | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exposure |   3.920884   1.241072     4.32   0.000     2.108406    7.291446
      gender |   2.566064    .807566     2.99   0.003     1.384797    4.754985
       _cons |    .196807   .0601514    -5.32   0.000     .1081141    .3582603
------------------------------------------------------------------------------
```

Non-Collapsibility

- Although sex is not a confounder, the effect estimates from univariable and multivariable regression analysis still differ.
- The difference of log(OR) is 0.067 and is entirely caused by noncollapsibility.
- Even in the absence of confounding, the univariable and multivariable exposure effect estimates might differ.
- The change-in-estimate based on logistic regression coefficients may lead to wrong conclusions when used to determine the presence of confounding.



Schuster, N.A., Twisk, J.W.R., ter Riet, G. et al. Noncollapsibility and its role in quantifying confounding bias in logistic regression. BMC Med Res Methodol 21, 136 (2021).

Randomized Controlled Trials

- Randomization ensures that all covariates will be balanced between treatment groups
- This is described as unconditional independence of treatment and prognosis (this means that treatment is given irrespectively of patient's characteristics that predict the course of the disease)
- Therefore the unadjusted difference in means and the adjusted difference in means will coincide
  - In RCTs, the estimated measure of treatment effect is **collapsible**.

Conditional and Marginal Effects

- Conditional effect: P(Y=1|A=1,X)-P(Y=1|A=0,X)
- It is adjusted for individuals' characteristics Contrasts the observed probability of outcome in treated and untreated individuals
- Can be generalised among people who actually received treatment
- A conditional treatment effect is the average effect of treatment on the individual.

Conditional and Marginal Effects

- Unconditional (unadjusted or marginal) effect: $P(Y=1|A=1)-P(Y=1|A=0)$
- Answers the question “what would have happened if all individuals had received the treatment”
- It can be generalised to the whole population

A marginal treatment effect is the average effect of treatment (ATE) on the population.

Why use a propensity score method

- It is an alternative method to adjust for confounding
- In some cases, as explained previously, conditional and marginal effects will not coincide
- Researchers should decide which effect is of interest and analyze accordingly

Why use a propensity score method

- A **propensity score based method** should be used when one needs to estimate a measure of **treatment effect for the population** that is **non-collapsible** (i.e. conditional and marginal effects do not coincide)
- It is also used in cases where a collider exists, which must be somehow included in the analysis
  - More on this issue will be discussed in the causal inference lecture that follows

Adjustment by Conditioning vs. Propensity Score

- **Adjustment by conditioning** gives results that are interpreted as the average **effect of treatment** on the individuals **that actually received treatment**.
- Techniques based on **Propensity Score** attempt to estimate the **effect of a treatment**, policy, or other intervention **on the population**, by accounting for the covariates that predict receiving the treatment.

Radical of Propensity Score

- Suggests to model the treatment rather than the outcome
- Provides a way to summarize covariate information about treatment selection into a single number (scalar)

Definition

- **The propensity score is the probability of treatment assignment (A) conditional on observed baseline characteristics (X).**

- It allows to design and analyze an observational (nonrandomized) study so that it **mimics** some of the particular characteristics of **a randomized controlled trial**.

Why use a propensity score method

- To reduce or eliminate the effects of confounding when using observational data to estimate treatment effects.
- It allows the estimation of the *marginal* rather than the *conditional* treatment effect in settings where adjustment should not be performed, as previously discussed
- Should be used when we need to assess the effect of an intervention to the whole population, and not just the treated individuals

General Idea

- Pr(A=1|X) is some marginal probability (e.g., 55%)
- The idea is to compare units who, based solely on their observables, had very similar probabilities of being placed into treatment
- If conditional on X, two units have a similar probability of treatment, then we say they have similar propensity scores
- We then think that all the difference in the outcome variable is due to the treatment.
- **If we compare a unit in the treatment group to a control group unit with two similar propensity scores, then conditional on the propensity score, all remaining variation between these two is randomness in selection on observables**

So, how does it works?

- It is a balancing score
  - This means that, conditional on the propensity score, the distribution of observed baseline covariates will be similar between treated and untreated subjects.

Example

Suppose we want to estimate the effect of eye surgery on patients with problematic sight (ameloriated sight (Y=1) vs. Not ameliorated sight (Y=0))

- The decision to operate (A) depends on patients age and disease severity (X)
- These two covariates also affect the surgery's outcome

Example

- 1st approach: We could regress the outcome (Y) (ameloriated sight vs. Not ameliorated sight) on A and X and estimate Pr(Y| A, X)
- Then, compare the **observed outcome in the treated and the untreated individuals given their characteristics**
- Pr(Y=1| A=1, X) and Pr(Y=1| A=0, X)

Example

- 2nd approach: We could estimate the probability of A given X, assign it to each individual and then estimate Pr(Y) taking into account their different *propensity* to receive treatment
- Then, compare the **unconditional** probability of the outcome in the treated and untreated individual, using the propensity score to create balance between treated and untreated

$$\Pr(Y=1|A=1) \text{ and } \Pr(Y=1|A=0)$$

Example

- In the 2nd approach, X are not included in the outcome model

- It is a marginal model

Some caveats

- This is only relevant for selection on observables
- If you cannot write down a conditioning strategy such that conditioning on X will satisfy the backdoor criterion, then this is not the research design you choose
- You need to identify the confounders, X, that will block all back doors and you will need data on them

Step 1: Estimation of the Propensity Score

- The propensity score is most often estimated using a logistic regression model, in which treatment status is regressed on observed baseline characteristics.
- The estimated propensity score is the predicted probability of treatment derived from the fitted regression model.

Example

- Stage 1: Fit a logistic regression model for surgery, rather than patient's outcome: logit(A|X)
- Get the predicted probability
- The model's estimate of Pr(A=1|X) is the propensity score

How is it estimated

- Although logistic regression appears to be the most commonly used method for estimating the propensity score, the use of other methods has been investigated, too.

Variable Selection for the Propensity Score Model

Possible sets of variables for inclusion in the propensity score model include the following:

- all measured baseline covariates
- all baseline covariates that are associated with treatment assignment
- all covariates that affect the outcome (i.e., the potential confounders)
- all covariates that affect both treatment assignment and the outcome (i.e., the true confounders)

Variable Selection for the Propensity Score Model

Lack of consensus as to which variables to include in the propensity score model.

However, recalling that the propensity score is defined as the probability of treatment assignment:

$$Pr(A=1|X)$$

Leads to theoretical arguments in favor of the inclusion of only those variables that affect treatment assignment.

Variable Selection for the Propensity Score Model-Results of recent studies

- Variables that do not affect exposure but that affect the outcome should always be included in the propensity score model.
- Including variables that affect exposure but not the outcome will increase the variance of the estimated treatment effect without a concomitant reduction in bias.

- **Always include variables that are associated with the outcome**

Brookhart et al. (2006).

Variable Selection for the Propensity Score Model-Results of recent studies

- According to recent studies, when matching with respect to the propensity score was performed, any of the above mentioned sets of covariates resulted in all prognostically important variables being balanced between treated and untreated subjects
- When **only the potential confounders or only the true confounders** were included:
  - Imbalances between treated and untreated subjects only on those variables that affected treatment assignment but that were independent of the outcome.
  - A greater number of matched pairs
  - Estimates of a null treatment effect with lower mean squared error
  - **Thus no additional bias and estimates of treatment effect with greater precision.**

Austin, Grootendorst, & Anderson (2007)

Practical Considerations

- In practice, may be difficult to accurately classify baseline variables into the true confounders, those that only affect the outcome, those that only affect exposure, and those that affect neither treatment nor the outcome.
- Published literature may provide some guidance for identifying variables that affect the outcome.
- In many cases, most individual baseline characteristics likely affect both treatment assignment and the outcome.
- **Therefore, one can safely include all measured baseline characteristics in the propensity score model.**

Practical Considerations

Variables that may require greater investigation are policy-related variables or variables denoting different temporal periods.

For instance, in a study comparing the affect of an older treatment with that of a newer treatment, subjects who entered the study in an earlier period may be more likely to receive the older treatment, whereas subjects who entered the study in a later period may be more likely to receive the newer treatment.

Thus a variable denoting a temporal period would affect treatment assignment.

However, if the outcome was conditionally independent of temporal period, the inclusion of a variable denoting temporal period in the propensity score model could result in the formation of fewer matched pairs compared with if this variable were excluded from the propensity score model

Caution! the propensity score model should only include variables that are measured at baseline and not post-baseline covariates that may be influenced or modified by the treatment.

See for example: The examination of the effect of atypical vs. typical neuroleptic agents on death in elderly nursing home residents with dementia; Austin, Grootendorst, & Anderson (2007)

Step 2: Balancing

- The reason we estimated the propensity score, was to use it in a way to provide balance of baseline characteristics between treated and untreated individuals
- This can be done in more than one ways

Propensity score methods

- Matching on the propensity score
- Stratification on the propensity score
- Inverse probability of treatment weighting using the propensity score
- Covariate adjustment using the propensity score.

Example (cont.)

- Stage 2: Having estimated the individual's propensity score, assign it.
  - Practically, this means to construct a new variable in the dataset, that will contains individuals propensity scores
- Then, use one of the previously mentioned methods to balance treated and untreated individuals with respect to confounders X

Matching (1)

- Propensity score matching entails forming matched sets of treated and untreated subjects who share a similar value of the propensity score (Rosenbaum & Rubin, 1983a, 1985).

Matching (2)

- The most common implementation of propensity score matching is one-to-one or pair matching, using the nearest neighbor matching
- Can match one-to-many as well
  - Pairs of treated and untreated subjects are formed, such that matched subjects have similar values of the propensity score

Nearest neighbor matching

- Pairs an individual with a given propensity score with another with ‘closest’ propensity score.
- Often used algorithms
  - “greedy”: It goes through the potential matches and selects the closest unmatched option to match each time
  - “optimal matching”: Minimizes global balance over all matches
- Can be performed either with or without replacement

Greedy nearest neighbor matching- Example

1. Choose the treated individual with the highest propensity score
2. Select an untreated individual with the closest propensity score to the person picked in Step 1.
3. Choose a second treatment group member (in this example, with the next highest propensity score rank), match the second participant.
4. Repeat the process until all participants are matched.

Matching (3)

- Once a matched sample has been formed, the treatment effect can be estimated by directly comparing outcomes between treated and untreated subjects in the matched sample.
  - Reporting of treatment effects can be done in same metrics as are commonly used in RCTs.

Matching (4)

- Once the effect of treatment has been estimated in the propensity score matched sample, the variance of the estimated treatment effect and its statistical significance can be estimated.

Matching (5)

- Treated and untreated subjects within the same matched set have similar values of the propensity score.
  - Their observed baseline covariates come from the same multivariate distribution.
- In the presence of confounding, baseline covariates are related to outcomes.
  - Matched subjects are more likely to have similar outcomes than are randomly selected subjects.
- The lack of independence in the matched sample should be accounted for when estimating the variance of the treatment effect

Matching (6)

- In Stata, the psmatch2 module or the teffects can be used for propensity score matching.

- `psmatch2 depvar [indepvars] [if exp] [in range] , [outcome(varlist) pscore(varname) ai(integer k>1) mahalanobis(varlist) caliper(real) noreplacement descending common trim(real) odds index logit ties warnings quietly ate]`
- `teffects psmatch (outcome) (treatment covariates)`

Stratification (1)

- Stratification on the propensity score involves stratifying subjects into mutually exclusive subsets based on their estimated propensity score.
-

Stratification (2)

- Subjects are ranked according to their estimated propensity score.
- Subjects are then stratified into subsets based on previously defined thresholds of the estimated propensity score.
- A common approach is to divide subjects into five equal-size groups using the quintiles of the estimated propensity score.

Stratification (3)

- Within each propensity score stratum, treated and untreated subjects will have roughly similar values of the propensity score.
- Therefore, when the propensity score has been correctly specified, the distribution of measured baseline covariates will be approximately similar between treated and untreated subjects within the same stratum.

Stratification (4)

- Stratification on the propensity can be conceptualized as a meta-analysis of a set of quasi-RCTs.
  - Within each stratum, the effect of treatment on outcomes can be estimated by comparing outcomes directly between treated and untreated subjects.
- The stratum-specific estimates of treatment effect can then be pooled across stratum to estimate an overall treatment effect.

Stratification (5)

- Stratum-specific estimates are weighted by the proportion of subjects in each stratum.
- When the sample is stratified into K equal-size strata, stratum-specific weights of 1/K are commonly used when pooling the stratum-specific treatment effects, allowing one to estimate the ATE
- The use of stratum-specific weights allows the estimation of the Average Treatment Effect for the Treated (ATT)
- A pooled estimate of the variance of the estimated treatment effect can be obtained by pooling the variances of the stratum-specific treatment effects.

(Imbens, 2004).

Inverse Probability Weighting

- Inverse probability of treatment weighting (IPTW) using the propensity score uses weights based on the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment.
- The use of IPTW is similar to the use of survey sampling weights that are used to weight survey samples so that they are representative of specific populations

Inverse Probability Weighting

- A subject's weight is equal to the inverse of the probability of receiving the treatment that the subject actually received.
- Inverse probability of treatment weighting was first proposed by Rosenbaum (1987a) as a form of model-based direct standardization.

Covariate Adjustment

- Using this approach, the outcome variable is regressed on an indicator variable denoting treatment status and the estimated propensity score.
- The choice of regression model would depend on the nature of the outcome. For continuous outcomes, a linear model would be chosen; for dichotomous outcomes, a logistic regression model may be selected.

Covariate Adjustment

- The effect of treatment is determined using the estimated regression coefficient from the fitted regression model. For a linear model, the treatment effect is an adjusted difference in means, whereas for a logistic model it is an adjusted odds ratio.
- Of the four propensity score methods, this is the only one that requires that a regression model relating the outcome to treatment status and a covariate (the propensity score) be specified.

Covariate Adjustment

Furthermore, this method assumes that the nature of the relationship between the propensity score and the outcome has been correctly modeled.

Step 3: Checking Balancing

Quantitative balance checks:

Standardized mean differences (SMD) and variance ratios (VR) can be examined to see how much imbalance is in each covariate.

Formal statistical tests include the Kolmogorov–Smirnov (K-S) test or regression analysis where the outcome is each covariate and the treatment indicator is used as a predictor.

3. Checking Balancing

Graphical methods: Examine propensity score distribution plots or boxplots for each treatment group.

More overlap means better balance across the treatment groups.

Step 4: Outcome Anlaysis

After balancing across the treatment groups, outcome analysis can be conducted to estimate an average causal effect.

Average treatment effect (ATE): What is the difference between the treatment groups in the population?

Average treatment effect among the treated (ATT): What is the expected difference in the potential outcome when the control treatment was applied instead of the alternative treatment?

Propensity score matched samples and RCTs

Directly compare outcomes between treated and untreated subjects in an RCT.

On average, similar distribution of covariates between treatment groups.

But.. residual differences in covariates may exist between treatment groups.

Regression adjustment can be used to reduce bias due to residual confounding.

Regression adjustment increases precision for continuous outcomes and statistical power for continuous, binary, and time-to-event outcomes

(Steyerberg, 2009)

Directly compare outcomes between treated and untreated subjects within the propensity score matched sample.

Covariate balance is a large sample property, that can be violated in practice

Propensity score matching can be combined with additional matching on prognostic factors or regression adjustment

(Imbens, 2004; Rubin & Thomas, 2000)

Differences and Similarities of Propensity Score Methods

Propensity score matching, stratification on the propensity score, and IPTW differ from covariate adjustment using the propensity score:

The three former methods separate the design of the study from the analysis of the study

this separation does not occur when covariate adjustment using the propensity score is used.

Appropriate diagnostics exist for each of the four propensity score methods to assess whether the propensity score model has been adequately specified.

However, with propensity score matching, stratification on the propensity score, and IPTW, once one is satisfied with the specification of the propensity score model, one can directly estimate the effect of treatment on outcomes in the matched, stratified, or weighted sample.

Specification of a regression model relating the outcome to treatment is not necessary.

Differences and Similarities of Propensity Score Methods

When using covariate adjustment using the propensity score, once one is satisfied that the propensity score model has been adequately specified, one must fit a regression model relating the outcome to an indicator variable denoting treatment status and to the propensity score.

In specifying the regression model, one must correctly model the relationship between the propensity score and the outcome (e.g., specifying whether the relationship is linear or nonlinear).

In doing so, the outcome is always in sight because the outcome model contains both the propensity score and the outcome.

As Rubin (2001) notes, when using regression modeling, the temptation to work toward the desired or anticipated result is always present.

Another difference between the four propensity score approaches is that covariate adjustment using the propensity score and IPTW may be more sensitive to whether the propensity score has been accurately estimated (Rubin, 2004).

The Four Steps in Propensity Score Analysis

Steps	Procedure	Considerations
1. Propensity Score Estimation	Include baseline covariates Use appropriate model according to the type of the treatment variable	Variables' selection for the treatment model
2. Balancing	Use matching, stratification, IPW or adjustment by the PS	Matching can be infeasible/poor. Stratification can be difficult to implement. Weighting do not perform well when PS takes values close to zero or one.

The Four Steps in Propensity Score Analysis

Steps	Procedure	Considerations
3. Checking Balancing	PS distribution plots or boxplots for each treatment group. More overlap means better balance. Standardized mean differences (SMD) and variance ratios (VR) can be examined to see how much imbalance is in each covariate. Formal statistical tests include the Kolmogorov–Smirnov (K-S) test or regression analysis where the outcome is each covariate and the treatment indicator is used as a predictor	Criterion values for SMD and VR (e.g., |SMD| > .25 and VR > 2 or VR < 0.5) are arbitrary. Failure of balance might lead to re-estimating the propensity scores with other methods or adding in multiple variables in the estimation model. This might cause problems where a researcher utilizes multiple methods until they arrive at the desired conclusion

The Four Steps in Propensity Score Analysis

Steps	Procedure	Considerations
4. Outcome Analysis	After balancing across the treatment groups, outcome analysis can be conducted to estimate an average causal effect. Average treatment effect (ATE), Average treatment effect among the treated (ATT)	Misspecifications in the propensity score estimation model and/or in the outcome model can lead to erroneous conclusions. Doubly robust methods can be used to adjust for some of these misspecifications. Sensitivity analysis is also recommended to see how sensitive the results are to any model misspecifications