Propensity Score Methods in Clinical Studies: Improving Causal Inference

Assoc. Prof. Cameron Hurst

ANFPP & MWRC
Charles Darwin University

20th October, 2023

What we will cover...

1. Motivation for PS methods
   - Causal inference
   - The problem of confounding
2. Propensity score methods
   - Concepts and rationale
   - Different PS methods
   - ATE or ATT?
   - Checking balance
   - PS adjustment
   - PS matching (and stratification)
   - PS weighting
3. Sensitivity analysis and other issues
   - Extending basic PS analyses

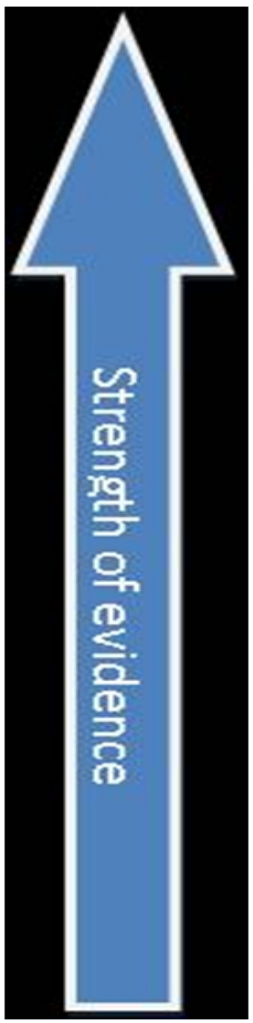
Causal inference in the clinical sciences

- When conducting clinical studies, particularly ones involving highly refined research questions (i.e. hypothesis testing), the ability to infer **causal relationships** (causality) is of the utmost importance.
- With this in mind, a set of **study designs** have been developed.
- Although, we try to choose this study design (stronger evidence), due to practical reasons it is more often that the study design chooses us.

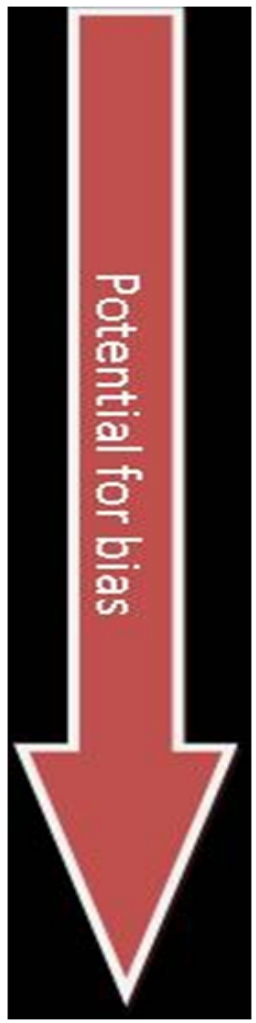
Causation vs association

While association is a pre-condition of causality (it's *necessary*), association alone is not *sufficient* to conclude causality

Causal inference: Study design and strength of evidence

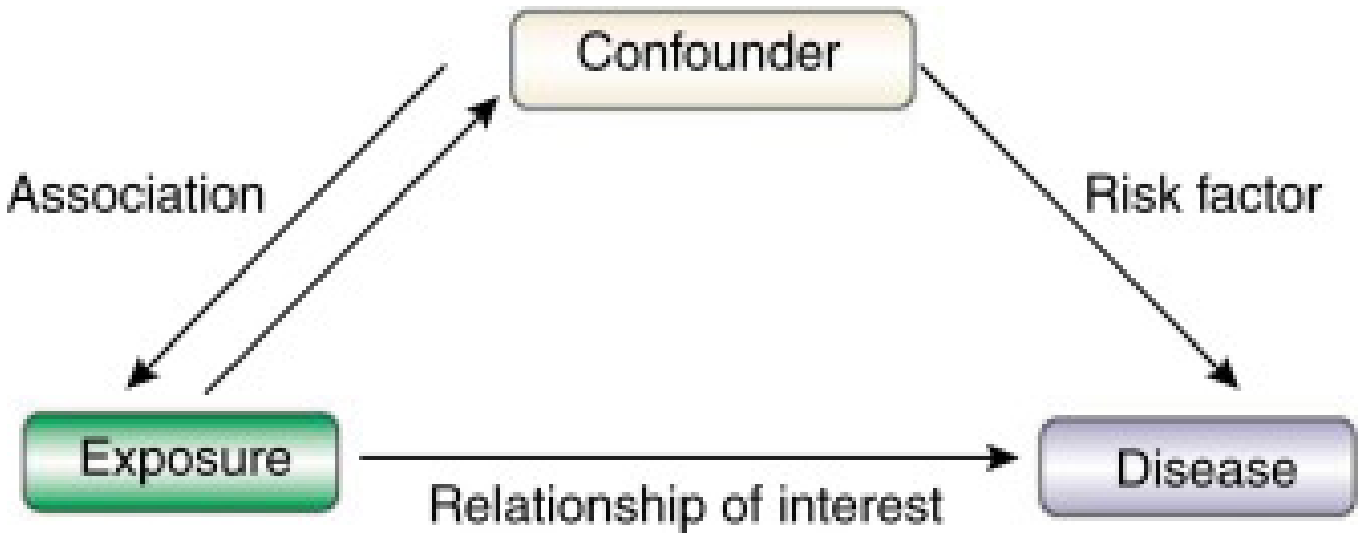


Study design
Meta analysis
Randomized Controlled Trial
Quasi-experiment
Cohort study
Cross-sectional
Case-Control
Ecological
Case series
Case study



Study design and the problem of confounding

- There are many types of bias that can cause problems with our studies
- One of the most important, and hardest to deal (adequately) with, is **Confounding bias**



Confounding: A toy example

Let's say we are interested in a new drug to treat hypertension

- Outcome: SBP $\geq$ 130 mm Hg AND/OR DBP $\geq$ 80 mm Hg
- One arm takes tablet containing *Drug-X* once a day, and the placebo-controlled arm takes a similar looking capsule with no active ingredient.
- We run a **univariate** log-binomial regression and get Relative Risk of 0.75 ($RR = 0.75$, $95\% CI : 0.65, 0.85$, $p < 0.05$) .
  - Those in the drug arm have 0.75 the risk of elevated BP relative to patients in the control arm.

All good?

The problem of confounding

- My framing of the problem impied this study was a trial (an experimental study), but I didn't mention randomization.
- In fact in this case it was a quasi-experiment (no randomization), and we find out that our control group is, on average, 83 years old, vs an average age of 74 in our treatment arm.
- So, is it the drug, or the younger age, that explains lower risk of elevated BP; **Confounding**

The concept of confounding bias

Confounding is where we can't separate the effect of interest, from some other nuisance effect.

Dealing with the confounding

So how do we solve the problem of confounding? Broadly speaking, two main approaches:

1. **Study design:** The best way to deal with confounding is to 'design it out'. The easiest way to do this is through **randomization**. Randomization (for large sample sizes) generally ensures that potential confounders will be balanced among different experimental arms. **Control** can also reduce confounding (e.g. only considering specific substrata of the population), but tends to reduce the generalizablility of our study.
2. **Modeling:** We can statistically control for confounders by adding them to our model (one way or another). The most common way of doing this is through standard multivariable modeling.

Experimental design or statistical modelling?

- There is almost no argument among clinical researchers that experimental design is the best way to control for confounding (this is why the RCT is considered the gold standard) BUT there are very good reasons why we can't always use RCTs:
  1. **Effort:** It is often not feasible to run RCTs (Time, Effort, $$$)
  2. **Ethics:** There are many ethical hurdles to RCTs (part of the effort problem). Perhaps the most obvious is if we want to study risk factors (it is highly unethical to allocate people to different risk groups)
- So in reality, we are often left with statistical control (**ENTER: PROPENSITY SCORING**)

The idea of statistical control: HTN and Drug X

So back to our simple example. Our original analysis, had:

$$ElevBP = f(Treatment)$$

Which gave us $RR = 0.75$, $95\%CI : 0.65, 0.85$, $p < 0.05$
NOW if we adjust for age (using a multivariable model):

$$ElevBP = f(Treatment, Age)$$

We get $RR = 0.97$, $95\%CI : 0.92, 1.02$, $p < 0.75$
In other words our naive analysis was misleading, if we (statistically) make both groups the (sample) average age, we can **no longer conclude that our drug had a significant effect in reducing elevated BP**

'Modeling out' confounders

We are also not limited to a single confounder. For example:

$$ElevBP = f(Treatment, Age, Sex, Education, TimeSinceDiag, ...)$$

Such a model would give us a more 'pure' treatment effect after controlling for all of these different covariates.

How much is too much?

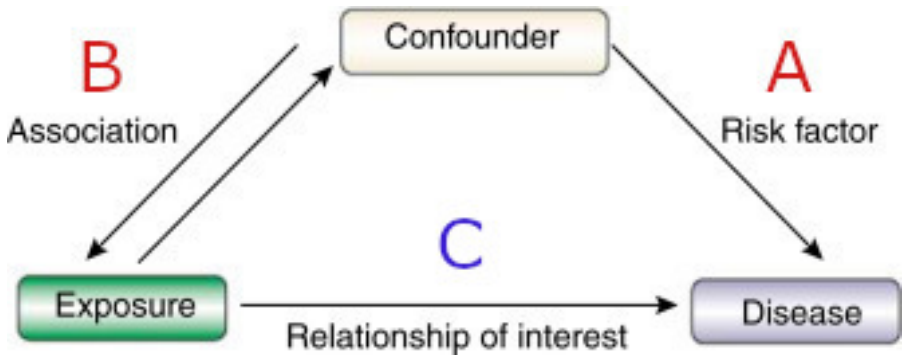
Consider:

- **Age** **not balanced** between groups, and associated with BP;
- **Sex/Gender** also **not balanced** between groups and is associated (somewhat) with BP;
- **Eye colour** (Brown, Blue, Green) was **not balanced** among arms, but is **NOT** associated with BP

Which are confounders? Which should be adjusted for?

To adjust or not to adjust? That is the question

- So given many covariates, some of them are:
  1. **Confounders** (associated with outcome AND effect of interest: **A** and **B**)
  2. **Independent risk factors** (Important risk factors, but do not confound with our effect of interest: **A** only)
  3. **NOT** associated with the outcome, but **ARE** associated with effect of interest (e.g. eye-colour): **B** only
  4. **Neither** associated with our effect of interest, NOR our outcome (NEITHER **A** nor **B**)
- For which should we adjust?



The idea of a propensity score

- Obvious that even 6 or 7 covariates will make our model unweildly, and 'readers' will become distracted
- Rosenbaum and Rubin came up with the idea of collapsing many confounders into a single variable, the **propensity score**
- It is called the **propensity score** because it represents the propensity (i.e. tendancy) for a patient to be in one group (e.g. treatment arm) compared to another treatment arm.

What is the propensity score?

The **Propensity score** is the **propensity**/**tendency**/**likelihood** a patient is in the treated group (based on their 'characteristics')

Rubin is the same statistician so important in the meta-analytic and multiple imputation area

The propensity score defined

The propensity score (for a individual patient) represents the **probability of that patient being in the treatment arm**, given their observable characteristics (Age, Sex etc). In other words,

$$P(X) = (T = 1|X)$$

What is a propensity score

In a well run 1:1 RCT the chance (**i.e. propensity**) for any and all patients to be in a particular treatment arm should be 0.5, AND it should be **independent** of that patient's characteristics

Propensity of a patient to be in an arm (a binary outcome) based on their characteristics is a **standard binary logistic regression problem** but also can be modeled using a **probit model**

Different ways of using the PS

Will will cover three ways of using propensity scores to try and offset confounding bias. These are:

- Regression adjustment (adding the PS as a model covariate)
- Propensity score matching (and stratification)
- Propensity score weighting (Inverse Probability Weighting)

Perhaps the best way to compare these various approaches is to run through a worked example (and then talk about each method's approach, advantages and disadvantages)

ATE or ATT?

Before we start our analysis, we need to choose our target population. Propensity score analysis provides two different effect estimates:

- **ATT** (**Average Treatment** effect in the **Treated**)
  The ATT (usually of interest to the clinical researcher.....us) is the effect of a treatment on **someone undergoing the treatment**. So what would I expect if I gave **MY PATIENT** this drug
- **ATE** (**Average Treatment Effect**): In contrast, if you were a Health economist or policy person (e.g. Worked for the MOPH), you may want to know if we funded this new treatment in Thailand, what effect might it have on the (whole) population burden; ATE is about the average effect on the **WHOLE POPULATION** (even those treated)

ATE vs ATT: Example

- ATE: Let's say the ATE is -10 mmHg. This means that, on average, the medication resulted in a 10 mmHg reduction in blood pressure across the entire population, which includes both those who took the medication and those who did not.
- ATT: Now, let's say the ATT is -15 mmHg. This means that, on average, **among the patients who actually took the medication**, the blood pressure reduced by 15 mmHg.

Effect size: ATE v ATT

- As ATE is an average between those treated, and those not, we would expect ATE $<$ ATT
- When using ATE, take care that our study (including ratio of treated to controls) represents how the drug to be ultimately administered in the clinical population (i.e. probably not 1:1 like in many RCTs)

Motivating example: Pregnancy outcomes and Asthma clinical practice guidelines

Journal of Asthma

Taylor & Francis
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ijas20

Do improvements in clinical practice guidelines alter pregnancy outcomes in asthmatic women? A single-center retrospective cohort study

J. L. Robinson, K. L. Gatford, C. P. Hurst, V. L. Clifton, J. L. Morrison & M. J. Stark

To cite this article: J. L. Robinson, K. L. Gatford, C. P. Hurst, V. L. Clifton, J. L. Morrison & M. J. Stark (2023) Do improvements in clinical practice guidelines alter pregnancy outcomes in asthmatic women? A single-center retrospective cohort study, Journal of Asthma, 60:10, 1907-1917, DOI: 10.1080/02770903.2023.2200824

To link to this article: https://doi.org/10.1080/02770903.2023.2200824

Example: Pregnancy outcomes and Asthma CP guidelines

- It is well established that (maternal) asthma can have a negative impact on perinatal outcomes
- In 2012, the South Australian government revised its clinical guideline to address management based on Asthma severity
- In total, $\sim$ 49,600 mother of which **3,711 mothers had asthma** $\sim 7.5\%$
- Many maternal and perinatal outcomes, but we will focus solely on the Small for gestational age binary outcome
- Overall study detailed (lots of research questions), but **TODAY** we want to compare SGA rates among babies with Asthmatic mums (vs those of non-asthmatic mums)

Problems?

Clearly an observational study. Any 'Quality of Evidence' issues?

ATE or ATT

In this asthma study:
ATE or ATT???

Are we more interested in those treated, or rather EXPOSED to asthma (specifically), or is the overall overall population burden of Asthma (in the South Australian population) on maternal and perinatal outcomes more of interest?

For me? ATT

A large majority of studies I work on (predominately clinical), are more interested in the ATT. That is, what is the treatment effect on those (specifically) undergoing the treatment?

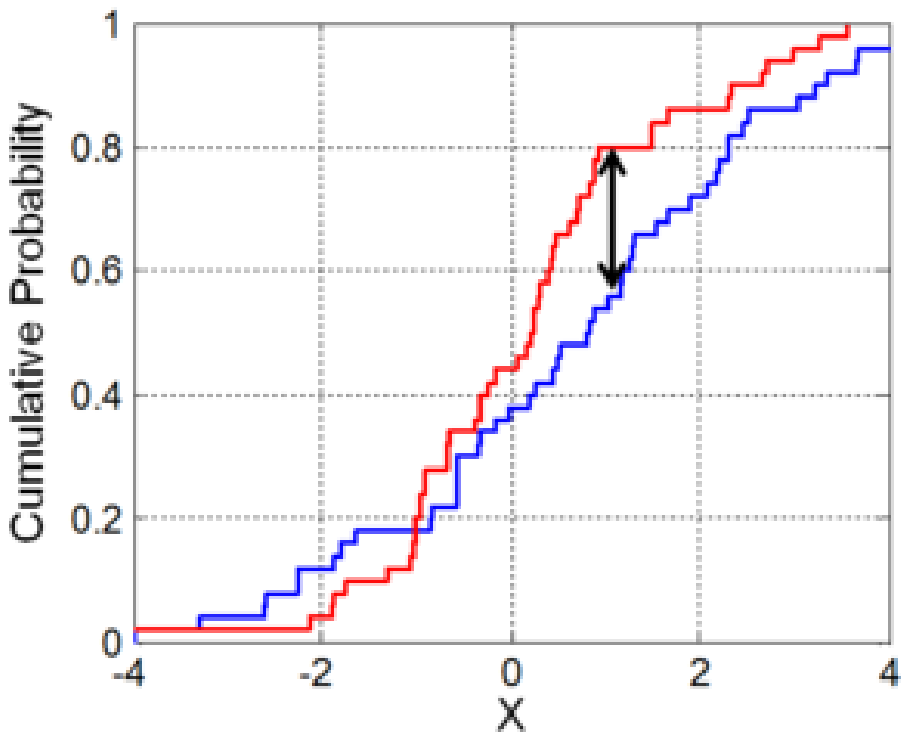
Comparing groups BEFORE adjustment(a subset)

Effect	Means.Treated	Means.Control	Std..Mean.Diff.	eCDF.Mean
distance	0.11002030	0.07201419	0.705837492	0.192134227
Epoch.01	0.60684452	0.53320832	0.150754560	0.073636196
AGE	29.18997575	30.00560377	-0.140465418	0.019907424
BMI.facHealthy	0.36998114	0.51168724	-0.293509372	0.141706098
BMI.facUnder	0.02209647	0.03429855	-0.083008838	0.012202078
BMI.facOver	0.26003773	0.25801317	0.004615373	0.002024556
BMI.facObese	0.34788467	0.19600105	0.318882798	0.151883621
Race_simplified.fac1	0.82295877	0.60043173	0.582983881	0.222527041

- **distance**: multivariable 'probit' difference between groups
- **Standardized mean difference** (Cohen's $d = \frac{meandiff}{sd}$): How many *sd*s apart the groups
- **eCDF** is the **Empirical Cumulative Distribution Function difference** (Smirnov statistic)

Aside: eCDF to compare two distributions

Most comparisons of two samples focus solely (or predominately) on its centre (its typical value). Comparing the Cumulative distributions of two samples **also** accounts for difference in variation, skewness and kurtosis.



Standard Multivariable adjustment

If we peruse the table (just a small subset):

Effect	Means.Treated	Means.Control	Std..Mean.Diff.	eCDF.Mean
distance	0.11002030	0.07201419	0.705837492	0.192134227
Epoch.01	0.60684452	0.53320832	0.150754560	0.073636196
AGE	29.18997575	30.00560377	-0.140465418	0.019907424
BMI.facHealthy	0.36998114	0.51168724	-0.293509372	0.141706098
BMI.facUnder	0.02209647	0.03429855	-0.083008838	0.012202078
BMI.facOver	0.26003773	0.25801317	0.004615373	0.002024556
BMI.facObese	0.34788467	0.19600105	0.318882798	0.151883621
Race_simplified.fac1	0.82295877	0.60043173	0.582983881	0.222527041

At least some difference $\Rightarrow$ lack of balance $\Rightarrow$ need to adjust

Assessing balance

Later I will show you a more succinct way to compare groups (but I wanted to introduce standardized mean difference and eCDFs)

Regression adjustment

Balancing by just adjusting for the propensity score (similar to just using multivariable adjustment) is simply running a standard bivariate model with the **addition of the PS score in the model**. That is:

$$SGA = f(Asthma, PScore)$$

In this case, this would be a standard binary logistic regression (or log-binomial regression) with:

- **SGA** - small for gestaional age)(OUTCOME)
- **Maternal Asthma** - Did mother (of baby) have Asthma (effect of interest)
- **PScore** a single covariate representing ALL confounders derived using Logistic or Probit regression

Regression adjustment: RESULTS

Now let's compare three pretty simple models:

1. Bivariate model (Crude effect)
2. Standard MV model (Adj effect)
3. Propensity score adjusted (as a covariate) model

Adjustment	OR	L95	U95	P.val
Crude (none)	1.328	1.214	1.453	0.00000
MV adjusted	1.180	1.073	1.297	0.00062
PS (reg) adjusted	1.175	1.072	1.288	0.00059
Propensity score	24.684	15.175	40.151	0.00000

It is noteworthy that the PS had a significant effect (SGA relates to the PS score confirming it is made up at least partially by confounders, but the OR itself is rather meaningless)

Interpreting results (so far)

- Crude model suggest that Maternal asthma substantially increases th risk of SGA ($OR = 1.323$, $95\% CI : 1.214, 1.453$, $p < 0.001$)
- However, when we adjust for our potential confounders (either approach), there was a substantial reduction in the magnitude of risk, albeit, still statistically significant ($p < 0.001$ for both models)
- Standard MV and Regression PS-adjusted model were similar.

Now, let's go onto to our more sophisticated Propensity score methods.

PS methods and the use of the PS

- All propensity score methods use the same propensity score.
- It is **HOW** these methods utilize the PS that differs:
  1. **Regression-adjustment** just adjusts for the PS by adding it as a model covariate
  2. **Matching** (and stratification) methods use the PS as a basis for matching (or stratification). **That is, patients with a similar propensity score are matched (or put in the same stratum)**
  3. **Inverse probability weighting** use the PS as a basis for WEIGHTING each patient (i.e. calculating each patients contribution to the overall estimate, and its variance)

Matching: The old-school matched case-control studies

Idea of matching been around a long time. In **Matched case-control studies**, participants with (cases) and without (controls) a disease would be selected and then **matched** on the basis of Sex, Age and Socio-economic status, Then **Conditional logistic regression** would be performed. Note:

- We can match anywhere from 1:1 (Case:Control) to 1:4. After this, NO gain to the bigger groups sample size; **diminishing return** with each additional control we add
- These matching groups represent a random effect $\Rightarrow$ **Conditional logistic regression similar to a random-intercept Generalized Linear Mixed Model**

Matched case:control studies and conditional logistic regression

- In matched case-control studies, Sex, Age and SES were known (& assumed) confounders. Conditional logistic regression was a forerunner of PS-matched models
- For matched propensity score models, all that we are really doing is extending our matching a little further by:
  1. Adding some more covariates (Confounders?)
  2. Differentially weighting these covariates, empirically, based on their association (level) with our effect of interest (X)

What do you think?

Remember a confounder is a variable **associated with BOTH the outcome AND the effect of interest**, are we using both of these associations to identify variables to add to our PS? More later.

Methods for matching

- Many different methods now used for matching based on Propensity score. For example, the R library I am using (`MatchIt`) has nine, several of them based on machine-learning algorithms
- The selection of methods to use depends on issues such as: ATT vs ATE; or do we want 1:1 or 1:k matching.
- I will use the **Nearest Neighbour** method (*aka* Greedy matching) with 1 (case) to 4 (controls) matching.
- Note that Nearest Neighbour not appropriate for ATE

1:1 vs 1:k matching

If we have $n_1 >> n_2$ then 1:1 matching will result in many observations being dropped from our analysis (another potential source of bias)

Results of matching

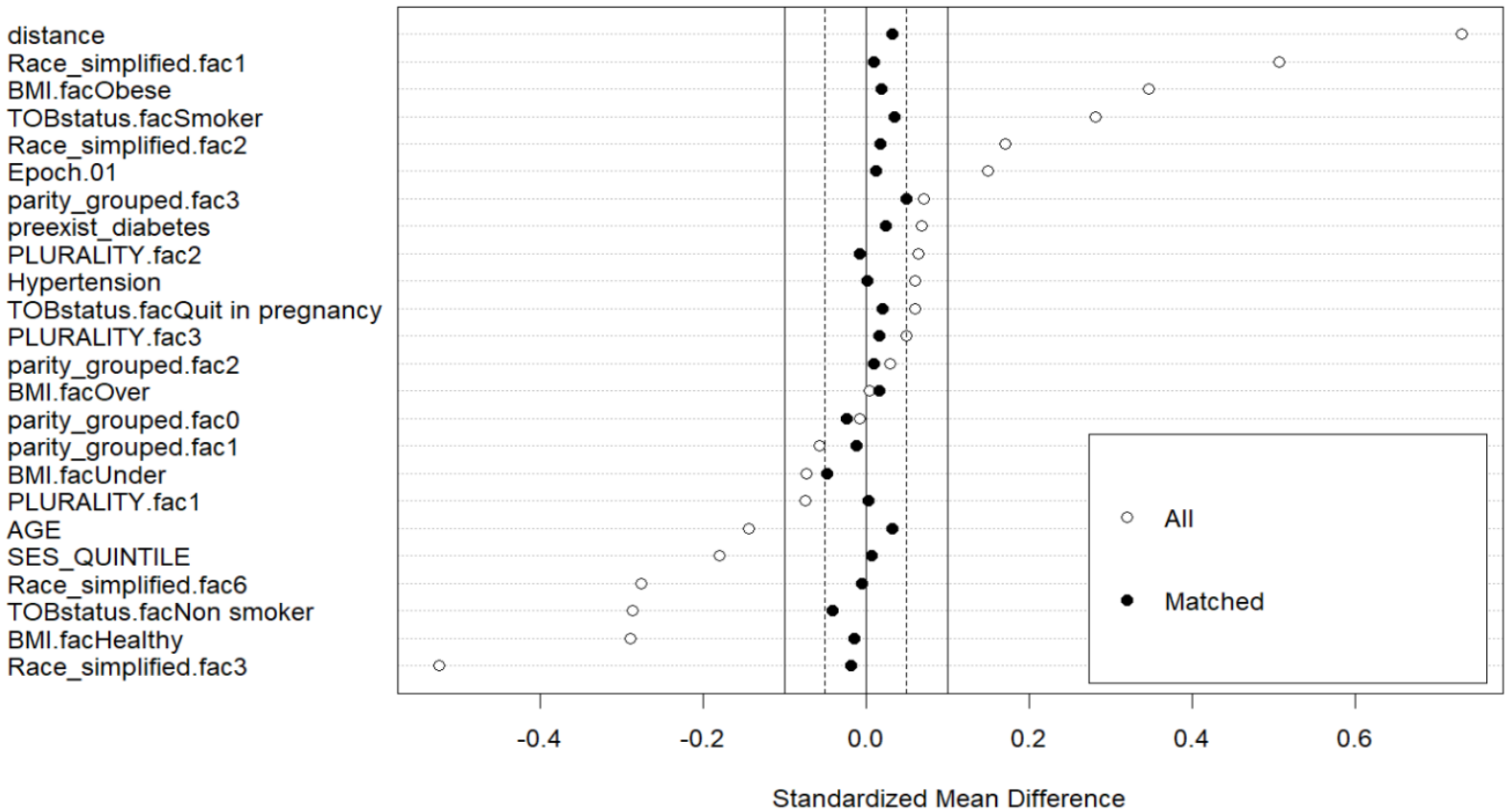
From my 'Nearest Neighbour' matching method, we get:

```
> m.out1.ATT
A matchit object
 - method: 4:1 nearest neighbor matching without replacement
 - distance: Propensity score
             - estimated with logistic regression
 - number of obs.: 49573 (original), 18555 (matched)
 - target estimand: ATT
 - covariates: Epoch.01, AGE, BMI.fac, Race_simplified.fac, SES_QUINTILE, TOBs
tatus.fac, parity_grouped.fac, preexist_diabetes, Hypertension, PLURALITY.fac
```

Even using 1:4 matching, the quite low prevalence of Asthma (7.5%) meant that we have dumped about 30,000 (over 60%) 'control' observations from our dataset.

Balancing the confounders: Does it work?

Has matching worked? Are our covariates better balanced post-matching (•) compared to pre-matching (○)?



Now to fit our PS-matched model

Adjustment	OR	L95	U95	P.val
Crude (none)	1.328	1.214	1.453	0
MV adjusted	1.18	1.073	1.297	0.00062
PS (reg) adjusted	1.175	1.072	1.288	0.00059
PS-Matched 1:4	1.17	1.065	1.286	0.0011

A couple of points to consider:

- The OR from the "matched" PS model weakened the effect marginally.
- So, is the slight weakening of the effect due to the PS-method (Matching 1:4), or the change in our sample (only about 40% of the original sample used)

The concept of 'calipers'

- Before leaving PS matching, sometimes (but not here) it's hard finding controls to match our treated/exposed patients
- In this case we can use the 'calipers' argument to **broaden** our matching criteria (e.g. a patient can be within 3 years of age, not the exact same age)
- We have had no problem finding matching controls in this study, our problem is a lack of exposed (treated) participants (only about 7.5% of the sample)

REM: The diminishing return in unbalanced analyses

Remember: After about a 1:4 match there is no real gain in adding extra 'controls'. The smaller of the two sample sizes ($n_1 << n_2$) becomes the limiting factor (not the overall sample size, N)

The idea of weighting (in general)

- The last PS method we will consider (and considered best practice) is Inverse Probability Weighting (IPW)
- It has its genesis in the statistical sub-discipline of **Survey Sampling**.
- In **survey sampling** we can Up-weigh (or down-weigh) particular observations to make our sample **better represent the population**.
- In contrast, in the PS context we can use IPW to re-weight observations so our treatment and control group confounder profiles are more similar

Inverse Probability Weighting

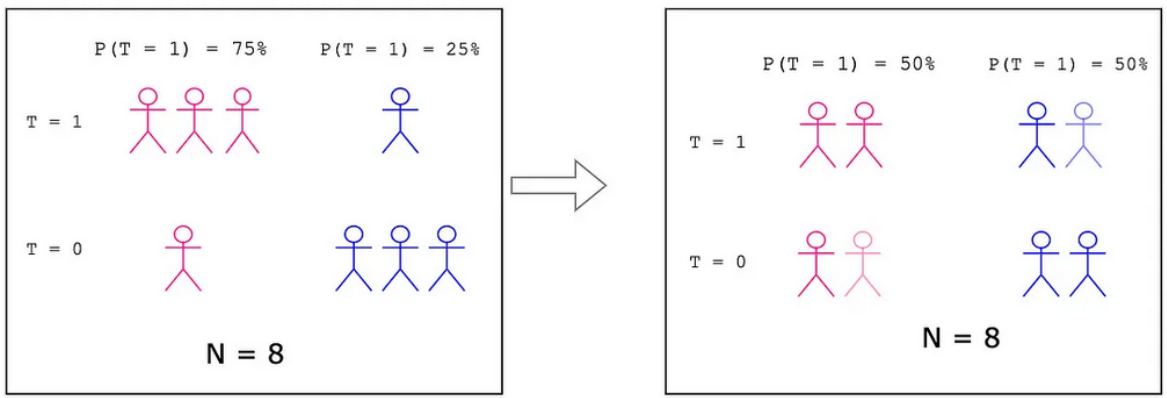
Inverse Probability Weighting (IPW), also known as Inverse Probability Treatment Weighting (IPTW), involves two steps:

1. Calculate the propensity of patients to be in a particular treatment groups (like every other PS method - ALREADY DONE)
2. Weight the individual patients by the inverse of the propensity score ($Weight_i = \frac{1}{ps_i}$).

A (very) simple example

A naive example

If we (for some weird reason) compared the heights of engineers and nurses, we would run into the problem that engineers are mainly men, whereas nurses are mainly women EG $P(F|T = NURSE) = 0.75$ and $P(F|T = ENGINEER) = 0.25$



We can use IPW to (artificially) equalize the probability of being female (or male) between the treatment (profession) groups.

PS Matching vs IPW

IPW has some distinct advantages:

- Matching is MAINLY limited to two treatment (or exposure) groups, **IPW can have more than two treatments**
- In fact, IPW can **deal with more complex designs** in general (e.g. longitudinal and other clustered designs)
- Much **simpler to implement**, and **runs easily on different models**, whereas matching methods have complex interplay between matching approach, estimation and caliper choices
- IPW **doesn't throw observations away** (recall we used only 40% of our observations when we used PS-matching)

Perhaps the main **disadvantage** of IPW (compared to matching) is IPW can be sensitive to: 1. **model selection** (adding/dropping a covariate can have a profound effect on results); and 2. **Extreme weights** (good idea to look at histogram of weights)

IPW: Results (compared to all methods)

Adjustment	OR	L95	U95	P.val
Crude (none)	1.328	1.214	1.453	0
MV adjusted	1.18	1.073	1.297	0.00062
PS (reg) adjusted	1.175	1.072	1.288	0.00059
PS-Matched 1:4	1.17	1.065	1.286	0.0011
PS-Inverse Prob. Weighting	1.189	1.078	1.311	0

- **Was all this PS stuff worth it?** Not really. Multivariable adjustment produced much the same results.
- **Should we use IPW or Matched-PS methods over standard MV adjustments?** YES!!! Especially, for Quasi-experimental studies (i.e. Non-randomized trials). Reviewers will expect you to use the more advanced PS methods (usually IPW).

Why PS methods (over standard multivariable modeling?)

In the end, **what are the advantage of PS methods over standard MV model?** Besides getting your manuscripts accepted, and sort of 'hiding your model complexity', there are no major advantages

Why do you think it is most important for Non-randomized studies?

Which covariates should be added to my PS-model

This is a very controversial issue. Even some 'experts' disagree on this. Let's (just for a couple of minutes) be democratic. I will give you three potential options, after which, I will ask for your vote.

1. **All confounders** - covariates associated with BOTH the outcome, Y, AND the treatment, X, (effect of interest) group
2. **All confounders** AND **All covariates associated with Y**
3. **All covariates in your dataset**

Did you think about ...

Now I want to introduce a few more issues (to see if you change your vote)

- Why are you doing this study? (Exploratory modeling, Predictive modeling or Estimation/Hypothesis testing)
- Occams Razor: The simplest solution is often the best (or in the statistical world: A simple model is often more generalizable and robust)
- Missing value bias (ie. Complete-case bias)
- Are we just limited to the effect of interest and the PS in our model? No. We can add additional covariates, in the usual way, and estimate their effect

Did anyone change their vote?

Sensitivity analysis

In the end, I run several combinations of PS methods and covariate selections to gauge the sensitivity of my estimates to these choices

I know we didn't really change our list of covariates, BUT do you think our results varied much across the different PS methods?

PS for non-trial analyses

A large majority of applications of PS methods (especially in the clinical domain) are for non-randomized trials or single-time point studies involving two arms. However (with a little work) PS methods can be extended to more advanced analyses. We should mote that:

- Both IPW and matching approaches can be extended to more advanced situations, BUT
- IPW lends itself MUCH better to more advanced application. Using matching for anything other than the simplest (2-arm, single time point) is a bit of a hassle.

I will only consider the longitudinal probelm here, but I suggest people interested in multi-arm studies (i.e. > 2 arms) see the TWAIN method (and library) in R.

PS in longitudinal (or otherwise clustered) studies

Perhaps the simplest way of using weighting (or matching) in a longitudinal study is the SPLIT-APPLY-COMBINE approach:

1. **SPLIT** your dataset into the different periods (e.g. three time points would give you three different datasets)
2. **APPLY** your Propensity score IPW or Matching approach to this single time point dataset and append the weight or matching variable to the dataset.
3. **COMBINE** your datasets back together.

Now If you are using matching, run your conditional model (e.g. Mixed model) using your matching variable* as the random effect. If you are using weighting, just specify your weighting variable as usual. **Only use this approach if focus is on group comparison, not comparing a group with its own baseline**

Software (and libraries)

- Anyone who knows anything about me will know I am an R fanatic (Have been using it for over 20 years).
- All of the analysis presented in this lecture was done using R, the `MatchIt` library for the Matched PS, and the `ipw` library for the Inverse Probability Weighting (which also requires the `survey` R library). I will provide my dataset and R script file to anyone who is interested.
- I have also used Stata for PS analyses (although this was years ago, and at that time I found it a bit cumbersome, BUT TO BE FAIR, that's probably because I am not really a Stata guy)

Questions

THANK YOU

QUESTIONS?