Propensity Scores in Clinical Research:

Laurel Beckett, PhD
Daniel Tancredi, PhD

University of California, Davis

July 14, 2014
Outline

Setting
- Ideal case
- Real world

Controlling for selection bias
- Regression adjustment
- Propensity scores

Examples
Propensity scores are a tool for helping to control for bias due to heterogeneity and imbalance in comparative clinical studies.

Propensity scores have been used extensively in studies in anesthesiology and critical care, but the quality of reporting of these studies needs improvement (Gayat 2010 PMID: 20689924).

The goal today is to provide an introduction and show some examples of propensity scores as used in anesthesiology and critical care studies.
You should be able to:

- Describe the study setting in which propensity scores are most helpful.
- Explain how propensity scores differ from regression adjustment.
- List several ways propensity scores may be used to correct for bias.
The ideal: direct comparison

Suppose a physician wants to compare Treatment A to Treatment B. The question for patient care is: would the patient’s outcome be better with A or B? Ideally, we would answer this question by comparing the two treatments directly in the same patients, and seeing which worked better.
Example of direct comparison

Sometimes you can actually do direct comparisons in the same patient:

- Topical treatments (left vs. right arm)
- Treatment for chronic pain (crossover design)

Example: high-flow O₂ vs. placebo (air) for treating cluster headaches.
Crossover design (Cohen, *JAMA* 2009 PMID: 19996400).
A total of 76 patients had both treatments.

- 24% or n=18 had pain relief at 30 min with air
- 72% or n=55 had pain relief at 30 min with O₂
- They don’t say how many had relief with both. Suppose it was 10 (14%).
- That means 8 had relief only with air and 45 only with O₂.
- This imbalance is very unlikely to be due to chance.

The study presents strong evidence that O₂ works better in most people.
Simulated data: differences in pain relief

This plot shows differences in pain relief in a hypothetical study where each patient tried both A and B. A lot of scatter here, but the mean level of pain relief is better with Treatment A (dashed line).
Simulated data: adding a confidence interval

A 95% confidence interval for the difference shows that the mean favors A more than chance alone.
You can’t always compare directly

What if you can’t compare directly?

- Sometimes you can only treat once (post-op pain).
- Sometimes you can only measure the outcome once (time to death).
- Then you have to compare a group of patients treated just with A to a group treated just with B.

The problem: Patients vary in how they respond to A and B! (Some patients actually respond to placebo but not to $O_2$, some to both, some to neither).
Approach 1: randomized clinical trials

If you can’t treat each person with both A and B, you can randomly assign to treatment groups.

- Every patient had same chance to get treatment A.
- Ensures that differences between groups are either due to treatment or to chance.
- We can calculate how likely it is that differences are just chance.
- Very unlikely that a post-op O$_2$ group would get all the easy-to-treat people and the placebo group all the hard-to-treat ones.
A key feature of randomized trials

With randomization, we can obtain unbiased estimates of the mean difference between outcome for a patient given A and the same patient given B.

Thus randomization makes up for our inability to test both treatments in the same patient.
You can’t always get what you want

You can’t always do a randomized clinical trial! It may be difficult or impossible:

- In emergencies
- If a control group would be unethical
- If informed consent is difficult

If you can’t do a randomized trial, you can still compare observational data, but with caution.
When treatment selection is not randomized

In observational studies, investigators do not control treatment assignment.

- You lose the protection given by randomization.
- Study arms may differ considerably:
  - in observed characteristics of patients
  - and also in unmeasured characteristics.
- Such imbalance can lead to *selection bias*. 
Selection bias

If certain kinds of patients are differentially assigned to Treatment A compared to B, we have selection bias.

Suppose, for example, pain patients in ER who had poor response to other treatment were more likely to be assigned to $O_2$.

Our estimate of the difference in outcome for Treatment A vs B would be biased.

What can we do?
Two approaches are commonly used to control for selection bias:

- Adjustment via regression modeling of outcome
- Assessing probability of different treatments and controlling via propensity scores.
Regression adjustment for selection bias

Basic assumptions in regression adjustment include that:

- the treatment groups differ in known prognostic factors,
- outcome is related to those factors, and
- there are not unmeasured confounders.

We then use regression models to estimate the average effect of treatment on outcome “conditional on” the values of the prognostic factors.
Regression models used for adjusting for bias

The type of regression model we use depends on the outcome variable.

- If the outcome is quantitative, like a pain scale, we usually use linear regression.
- If the outcome is dichotomous, like breakthrough pain, we use logistic regression.
- If the outcome is censored survival data, like time to needing surgery, we use proportional hazards models or other survival models.

We usually include prognostic factors through a linear combination of variables.
Interpretation of regression models

The regression model constructs a prognostic scale X (like the Apache score). We assume:

- Average outcome gets worse as you move from one end of scale to other.
- The observed difference in mean outcome between treatments for people who are at the same place on the prognostic scale is unbiased.
- The difference doesn’t depend on where you are on the prognostic scale.
Simulated data: regression model for pain relief

This plot shows association between prognostic score (X axis) and pain relief (Y axis). A lot of scatter but Treatment A seems a little better at all scores.
Simulated data: regression estimates shown

This plot shows regression lines for each treatment. Prognostic score predicts outcome. Treatment A has about 3 points greater relief at all scores.
The propensity score approach assumes that the treatment groups differ in known covariates, and develops a score $Z$ predicting how likely a person is to get Treatment A. We assume:

- The likelihood of getting treatment A is very similar for all people with the same value of the propensity score $Z$.
- People with the same propensity score $Z$ are like a little tiny randomized trial of A vs B.
Successful propensity scoring

In a randomized trial, the difference in prognostic factors between treatment groups is very small, on average.

A good propensity score would have small differences in prognostic factors, on average, if we compared just for people with similar propensity scores. Each little subgroup with similar scores would look similar, even if there is imbalance overall.

The propensity score also provides computational advantages over large multiple regression models in small sample size settings.
Effectiveness of correction for bias

Gayat shows a nice example of correction for selection bias.

Data are from a study on effect of epidural anesthesia on mortality after non-cardiac surgery.

Differences between treatments were much larger before propensity score correction (black circles) than after (red squares).
Correction for selection bias (Gayat)

Fig. 2 Graphical representation of absolute standardized differences before and after propensity score matching comparing covariate values. Example using Ref. [26]. Balance can be judged as successful as imbalance for all variables was less than 10%.
Using propensity scores

Four ways to use propensity scores to correct estimates of treatment effect for selection bias:

- Match patients on propensity scores
- Stratify patients on propensity scores
- Adjust for propensity score via regression modeling
- Estimate treatment differences using weights based on propensity scores
Propensity score weights refer to "potential outcome" distributions

In survey sampling, the basic weight for unit i is the inverse of being selected into the sample and refers back to the full population: \( w_i = 1/Pr\{i \in s\} \)

Similarly, the basic propensity score weight refers to either of the two populations of potential outcomes, according to observed treatment assignment:

\( w_i = 1/\pi_i \), if individual i received treatment

\( w_i = 1/(1 - \pi_i) \), if individual i received comparator
Two examples from the clinical literature

- One example from surgery literature
- One on anesthesia choice in prostatectomy
Impact of change of surgical teams

Brown *et al* (Annals Surg Feb 2011 PMID: 21173693) report a cohort analysis of non-emergent cardiovascular surgery outcomes. They compared:

- Day team (7 am to 3 pm)
- Evening team (3 pm to 11 pm)
- Changeover team (start in day, end in evening)

They used propensity scores to correct for selection bias.
Results from change of team study

Comparison of Day, Evening and Changeover teams:

- Unadjusted analyses showed major outcome differences.
- But patients differed in age, ejection fraction, comorbidity, type of procedure.
- After regression adjustment for propensity score:
  - C teams took 20+ min longer than D and E.
  - C had 1.85 fold greater risk of septicemia than D.

Explanation of propensity methods was skimpy!
Anesthetic choice for prostatectomy

Wuethrich et al compared outcomes for open radical prostatectomy using general anesthesia with or without epidural (Anesthesiology 2010 PMID: 20683253). Outcomes considered:

- Clinical progression-free survival
- Biochemical progression-free survival
- Cancer-specific survival
- Overall survival.
Analysis methods for prostatectomy study

The investigators used Cox proportional hazard models for all outcomes.

They used propensity scores because patients were not randomized to treatment.

- Distributions of scores too different in the groups to allow matching or stratification.
- One analysis used regression adjustment for propensity score.
- A second analysis adjusted by inverse probability weights.
Weighted results for prostatectomy study

PSA PFS: P=0.40

Clin PFS: P=0.002
A cautionary note from prostatectomy study

Although the overall conclusions were similar between the two methods used in this study, the point estimates and p values differed a fair amount.
For example, regression adjustment gave a hazard ratio for cancer-specific mortality of 0.95 (P=0.95) but inverse probability weights gave a hazard ratio of 0.45 (P=0.09).
Some concluding remarks

- Read the articles carefully; Gayat gives advice on reporting.
- Results may be sensitive to how propensity models are employed.
- Developing propensity scores is delicate and time consuming.
- Computer software for some aspects (matching, weighting) is not exactly push-button.

Despite the cautions, propensity scores remain a very useful tool for addressing selection bias problem, and far better than ignoring it!
Thank You!

Disclosures: Professor Beckett’s recent funding comes from UC Davis, NIH grants (various), and honoraria from talks at CSUS, University of Rochester, Modesto Community College, Albert Einstein School of Medicine and to Roche. She serves on external advisory boards for the Alzheimer’s Disease Centers of Albert Einstein, UC San Diego, and Washington University, and for the Alzheimer’s Prevention Initiative.

Prof. Tancredi’s recent funding comes from UC Davis and various research grants sponsored by the NIH, AHRQ, the CDC and the California Air Resources Board. In addition, he has received consulting fees and travel support from the International Scientific Association of Probiotics and Prebiotics.
References


