Methods for Dealing with Confounding in Observational Studies: Propensity Scores

Kristin Sainani, PhD
February 1, 2019
Propensity scores

- A data compression technique specifically for confounders
Why propensity scores?

- Observational treatment studies are limited by the lack of randomization.
  - Factors that affect treatment selection also affect outcomes (confounding by indication).
- Propensity scores aim to address this treatment selection bias.
Example 1: Ross procedure vs. mechanical valves

- Retrospective cohort study comparing two treatments for aortic valve replacement: Ross procedure (autograft) versus mechanical valve (with optimal anticoagulation therapy).
- Outcome: late survival
- Previous studies had suggested a survival advantage for Ross patients, but could not rule out bias due to patient selection.
  - Those selected for the Ross procedure tend to be younger and in better physical condition.

Treatment groups differ greatly!

Examples of imbalances, from Table 1:

**Table 1.** Baseline Characteristics: Unmatched Cohort

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Mechanical AVR (n=406)</th>
<th>Ross Procedure (n=918)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male gender, n (%)</td>
<td>310 (76.4)</td>
<td>691 (75.3)</td>
<td>0.672</td>
</tr>
<tr>
<td>Mean age at surgical intervention, y</td>
<td>49.5±10.3</td>
<td>41.6±11.0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cause, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rheumatic</td>
<td>23 (5.7)</td>
<td>37 (4.0)</td>
<td>0.054</td>
</tr>
<tr>
<td>Missing</td>
<td>58 (14.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calcified/degenerative</td>
<td>311 (76.6)</td>
<td>333 (36.3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Preoperative DM, n (%)</td>
<td>20 (4.9)</td>
<td>26 (2.8)</td>
<td>0.055</td>
</tr>
<tr>
<td>Preoperative hypertension, n (%)</td>
<td>161 (39.7)</td>
<td>245 (26.7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Concomitant CABG, n (%)</td>
<td>145 (35.7)</td>
<td>38 (4.1)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

AVR indicates aortic valve replacement; NYHA, New York Heart Association; DM, diabetes mellitus; LVEF, left ventricular ejection fraction; LVH, left ventricular hypertrophy; LVEDD, left ventricular end-diastolic diameter; LVESD, left ventricular end-systolic diameter; CABG, coronary artery bypass grafting; and MV, mitral valve.
Example 2: rehabilitation vs. no rehabilitation post-stroke

- Retrospective cohort study comparing rehabilitation in nursing homes versus no rehabilitation for stroke patients
  - Outcome: community discharge; functional status
  - Patients who receive rehabilitation have better outcomes than those who do not receive rehabilitation, but they are also less disabled, better insured, and have more social support at baseline.
    - Does the rehabilitation itself improve outcomes or would these patients have done well regardless?

Traditional ways to control for these confounders...

- Stratification
- Matching
- Statistical adjustment
Issues with these methods...

- Stratification
  - How can we stratify on so many confounders?

- Matching
  - How can we match on so many confounders?

- Statistical adjustment
  - Groups may simply be incomparable.
  - High chance of residual confounding.
  - Cannot control for 24 confounders when there are only 36 events (36 deaths)!
With propensity scores...

- Stratification
  - Easy to stratify on a single number!

- Matching
  - Easy to match on a single number!

- Statistical adjustment
  - Easy to adjust for a single number, even if event rate is low!
Propensity scores

- In a randomized trial, all participants have the same probability of receiving each treatment.

- In an observational study, participants vary in these probabilities.
  - Propensity scores estimate these probabilities for each individual, given their covariates (clinical, social, demographic characteristics).
Propensity scores

- The propensity score is the probability of receiving a treatment given one’s covariates:
  - \( P \) (treatment A / covariates)

- For example, a young patient in good physical condition might have a 70% chance of receiving a Ross procedure; an older patient in poor physical condition might have a 30% chance.
Propensity scores

- Propensity scores are estimated using logistic regression:
  \[ \text{Logit (treatment A)} = \text{intercept} + \text{covariate 1} + \text{covariate 2} + \text{covariate 3} + \text{covariate 4} \ldots \]

- Yields a predicted probability of treatment A for each individual.

- Reduces a large number of covariates to a single number (a probability).
Propensity scores

- Patients with similar propensity scores are comparable, even if they vary greatly in their underlying characteristics.

- If a patient with a 70% propensity score received the Ross procedure and another with a 70% propensity score received a mechanical valve, then, in theory, any difference in outcome can be attributed to the treatment rather than to patient selection.
Propensity scores vs. Randomization

- Matching or stratifying participants based on propensity scores yields treatment groups that are balanced with respect to measured covariates.
- Randomization yields treatment groups that are balanced with respect to measured and unmeasured covariates.
- Propensity scores do not eliminate unmeasured or residual confounding!
Building the propensity score model

- Logit (Ross Procedure) = intercept + preoperative LVESD + preoperative creatinin + age + concomitant CABG + mixed valve disease + gender + …

- Included 23 predictors
Building the propensity score model

- Logit (Rehabilitation) = intercept + age + medicaid insurance + vision score + mood score + activities of daily living score + use of assistive devices + ...
- They used 112 predictors!
Building the propensity score model, steps

1. Identify potential confounders (related both to treatment and outcome).
2. Impute missing data.
3. Build a non-parsimonious model, potentially including quadratics and interactions.
4. Stratify or match on the resulting propensity scores; assess the balance of covariates in the treatment groups.
5. If balance is poor, refit the model including additional confounders or higher order terms.
Building the propensity score model: missing data

- If an individual is missing one datapoint for one covariate, they will be omitted from the logistic regression.
- Must impute missing data!
- For example, in the rehabilitation study, 7 of the 112 predictors were missing values for 0.5% to 5.5% of the sample. The authors appropriately replaced these missing values with the mean values from the non-missing data.
Building the propensity score model: fit the model

- Normal rules of model building don’t apply!
  - Don’t worry about parsimony
  - Don’t worry about overfitting
  - Try multiple interactions and quadratic terms
- But… do consider omitting variables that are unrelated to outcomes (which cannot be confounders):
  - Simulations show that including these does not improve balance or reduce bias, but may make it harder to find matches
Building the propensity score model: evaluate the model

- The model is a success if it balances the treatment groups with respect to covariates!
- If balance is not achieved, refit the logistic model.
Evaluating the propensity score model

- Evaluate balance with “standardized differences” in lieu of p-value tests.
  - P-value tests depend highly on sample size.
  - A non-significant p-value does not guarantee that the groups are balanced.
- Standardized difference is the mean difference between groups expressed as the percent of 1 standard deviation.
- Standardized differences < 10% are considered balanced.
Standardized differences

- Standardized difference = \[ \frac{\bar{x}_1 - \bar{x}_2}{\frac{s_1 + s_2}{2}} \times 100 \]

Example, Ross study, unmatched cohort:
- Mean (SD) of age for Ross patients = 41.6 (11.0) years
- Mean (SD) of age for mechanical valve patients = 49.5 (10.3) years
- Average SD = 10.65 years
- Standardized difference = \[ \frac{49.5 - 41.6}{10.65} \times 100 = 75\% \]
Love plots for absolute standardized differences for baseline covariates between patients with mechanical valve and patients with the Ross procedure, before and after propensity score matching.

Preoperative LVESD
Preoperative LVEDD
Preoperative creatinin μmol/L
Mean age at surg. intervention (y)
Concomitant reconstruction
Concomitant CABG
Previous Aortic Valve Operation
Previous Cardiac Operation
Preoperative LVH
EF > 50
EF < 49
Preoperative Lung Disease
Preoperative Hypertension
Preoperative DM
Rhythm: other than sinus
NYHA III / IV
NYHA I / II
Mixed Valve Disease
Aortic Valve Regurgitation
Aortic Valve Stenosis
Active Endocarditis
Calcified/Degener. Valve Disease
Rheumatic Valve Disease
Sex (male)

Absolute standardized differences (%)
Use of propensity scores: Evaluating overlap

Comparing the distributions of propensity scores in different treatment groups may reveal:

- Certain subjects for whom there are no good comparators. These subjects should be excluded from the analysis.

Evaluating overlap

- Propensity score distributions may reveal when patient populations are too divergent to make meaningful comparisons, e.g.:

Use of propensity scores

1. Stratification
2. Matching
3. Statistical adjustment
1. Propensity score stratification

- Stratify on propensity score.
- E.g., in the rehabilitation study, authors stratified on quintiles of propensity score.
- Data are analyzed with Mantel-Haenszel methods for stratified data.

Their analysis revealed a significant interaction between propensity score and treatment effect!

The summary relative rate is 1.58, but the relative rate varies significantly by propensity for treatment.
2. Propensity score matching

- Matching on the propensity score optimizes matching.
- Many algorithms for propensity score matching:
  - Nearest-neighbor
  - Caliper matching
  - Mahalanobis metric matching in conjunction with propensity score
  - Others…
Propensity score matching

- Nearest-neighbor method:
  - In the Ross study, authors randomly ordered the mechanical valve patients and then sequentially matched each one to the Ross patient with the closest propensity score. If no Ross patients had a propensity score within 25%, the patient was left unmatched and excluded.
  - In fact, matches could be found only for 253 of 406 mechanical valve patients.
Results, matched cohort (n=253 pairs)

Cumulative Survival

Log rank test: p=0.29

Survival at 7 years:
- Mechanical AVR: 97%
- Ross procedure: 95%

# Patients at risk
- M. AVR: 252, 251, 251, 250, 226, 176, 95, 21
- Ross: 236, 206, 177, 147, 114, 95, 74, 50


Copyright © American Heart Association
### Table 3.
Association of Procedure With Late Mortality

<table>
<thead>
<tr>
<th></th>
<th>Mechanical Valve</th>
<th>Ross Procedure</th>
<th>Hazard Ratio (95% Confidence Interval)</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>After matching, n</td>
<td>253</td>
<td>253</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>All-cause mortality</strong></td>
<td>5/1682</td>
<td>7/1310</td>
<td>1.86 (0.58–5.91)</td>
<td>0.29</td>
</tr>
<tr>
<td>Valve-related mortality</td>
<td>0</td>
<td>4/1310</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non–valve-related cardiac mortality</td>
<td>3/1682</td>
<td>1/1310</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non–valve-related noncardiac mortality</td>
<td>1/1682</td>
<td>2/1310</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>1/1682</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Propensity score matching

- Tradeoff between inexact matching and incomplete matching.
  - Inexact matching increases residual confounding.
  - Incomplete matching decreases statistical power and generalizability of results.
Propensity score matching

- Matched data may be correlated and should be analyzed as matched pairs.
- There is some ongoing debate on this issue.
3. Statistical adjustment with propensity scores

- Outcome = intercept + treatment + propensity scores (+ other covariates?)
Assumptions!

- Assumes a certain relationship between the propensity score and outcome (e.g., linear in the logit)
- Assumes no interaction between propensity score and treatment (unless you add an interaction term between PS and tx).
Statistical adjustment with propensity scores

- Is similar to adjusting for all the covariates used to calculate the propensity score.
- But...is beneficial compared with traditional adjustment when the ratio of events: confounders <10
Cox regression for mortality, Ross study (unmatched cohort)

- $\ln (\text{rate of death}) = \text{Ross (vs. mechanical valve)} + \text{propensity score}$

- $HR = 3.64 \ (95\% \ CI: \ 1.22 - 10.88)$

- Could be driven by extreme skewness of the propensity scores in the Ross group...
Ways to address unmeasured confounding...

- Propensity score calibration
  - Collect more detailed confounder information in a subset of the sample.
  - Use this information to adjust or “calibrate” the propensity score estimates in the full set of data.
  - Use the corrected, or calibrated, propensity score for analyses of outcomes.
Summary: Advantages of propensity scores

- Focus the researcher on the problem of confounding by indication.
- Reduce a large set of confounders to a single, intuitive variable.
- Reveal subjects that cannot be compared or instances when whole groups cannot be compared.
- Make statistical adjustment possible when the number of confounders is large relative to the number of outcome events.
Summary: Disadvantages of propensity scores

- Are inferior to randomization.
- Do not solve the problems of residual and unmeasured confounding.
- May give the researcher/reader a false sense of security.
- Offer little benefit over traditional statistical adjustment when the ratio of outcomes/sample size:confounders is large.
- Are the subject of ongoing statistical debate, e.g.: “Why Propensity Scores Should Not Be Used for Matching” King and Neilson, November 2018 preprint.