The essentials of propensity score methods:
An application to assess the effect of inotropic therapy in cardiac surgery

Lehana Thabane, PhD
Professor, Clinical Epidemiology and Biostatistics

Department of Anesthesia Research Meeting
An RCT is the most optimal design for evaluating causal links.
The primary goal is prevention of bias:
Inaccurate measure of the Tx effect
What type of bias?

- Randomization ✓
  - Selection bias ✓

- Blinding ✓
  - Performance bias (for patients/providers) ✓
  - Detection, information or ascertainment bias (for outcome assessors) ✓

- Allocation concealment ✓
  - Selection bias ✓

- ITT analysis ✓
  - Attrition bias ✓

- Reporting Completeness ✓
  - Reporting bias ✓
The importance of RCTs

(Schulz KF. Randomized Controlled Trials: Introduction to Sequence Generation)

✓ In theory, RCTs are the gold standard in generating evidence to inform EBP

✓ In practice, they are frequently a bronze standard

Our challenge is to preserve the gold standard status of RCTs = Quality evidence to inform practice + policy
Analysis should match the objective and corresponding outcome

- **Binary outcome** (mortality/death)
  - **Yes**
  - Tests: Chisquared/Fisher's exact test, Logistic regression
  - Report: OR/RR/RD (95%CI);p

- **Time-to-event outcome** (time-to-death)
  - **Yes**
  - Tests: Log-rank test, Survival analysis/Cox-regression
  - Report: Hazard ratio (95%CI);p

- **Continuous** (SBP/DBP)
  - **Yes**
  - Tests: T-test/ANOVA
  - Report: Mean Diff (95% CI);p

- **Counts** (number of hospitalizations)
  - **Yes**
  - Tests: Poisson regression, Negative binomial regression
  - Report: Incidence Rate Ratio (95% CI);p
We often rely on observational studies when

- RCTs are too costly, unethical or not feasible
- Investigate relationships between exposures and outcomes in real practice settings
Observational Studies present major challenges

- Differential propensity to receive drugs/treatment *(channeling bias)*
  - It is a form of *allocation bias*
  - Preferential prescription of drugs on the basis of varying patient prognostic or provider characteristics

- Imbalances in risk/prognostic factors between groups

- Mixing of effects of risk/prognostic factors and the treatment/exposure *(confounding)*

- Biased estimates of the treatment effects
RCT

Randomization

No Randomization

Non-RCT
How is channeling bias or confounding handled in observational studies?
Commonly used methods

Population restriction

Regression

Stratification

Matching
Propensity score method: the what and how
The propensity score is the probability of treatment assignment conditional on observed baseline characteristics.

In RCTs the true propensity score is known from the study design.

In observational studies, the propensity score must be estimated using the sample data.
The propensity score is a balancing score.
The primary aim of propensity score methods is to mimic some of the characteristics of an RCT

- To compare outcomes in patients who have a similar distribution of (observed) baseline characteristics

- By comparing treated and untreated subjects with the similar propensity score, one can reduce or minimize observed confounding.
How is propensity score used to achieve balance?
PS Matching

Est PS  I  vs.  C

PS1

PS2

PSm

✓ Conditional logistic regression on matched pairs
✓ 1: M matching is possible
✓ Problem: may exclude unmatched patients
Stratified on PS

I vs. C

Est Rx effect

1st subclass

2nd subclass

5th subclass

◎ Patients **stratified** to according to estimated PS
◎ Patients within a subclass are more **homogeneous**
◎ Estimate Rx effect using a logistic regression within each subclass
◎ **Pooled estimate** = \((\theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5)/5\)
PS Weighting

- Inverse probability weighting in conjunction with regression modeling

- Patients in different Rx groups are re-weighted to represent the study population

- For patient receiving intervention
  \[ \text{weight} = \frac{1}{PS(X)} \]

- For patient receiving control
  \[ \text{weight} = \frac{1}{1 - PS(X)} \]
Inverse PSW Method

Propensity score  = 0.10

\[ IPW = \frac{1}{0.10} = 10 \]

= 9 replicates
Covariate Adjustment using PS

- For a binary outcome, include est. PS as a covariate in multiple logistic regression

\[
\text{logit}(\Pr(\text{event}|\text{covariates},Rx)) = B0 + B1*\text{PS} + B2*I(\text{Rx Group})
\]

- May also include quadratic terms or smoothing function for PS
How to correctly perform a propensity score analysis

1. Estimate Propensity score
2. Assess balance in baseline covariates between treated and non-treated subjects
3. Compare outcomes
   - If unacceptable balance, repeat steps 1 and 2 until acceptable balance is achieved.
Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples

Peter C. Austin¹, ², ³, *, †

Measures of balance should not be influenced by sample size and should be properties of the sample

- Standardized differences
- Variance ratios
- Empirical QQ plots
- Density curves
- Empirical cumulative distribution function (CDF)
Standardized differences

For continuous variables

\[ d = \frac{100 \times (\bar{x}_{\text{treat}} - \bar{x}_{\text{control}})}{\sqrt{\left(s^2_{\text{treat}} + s^2_{\text{control}}\right) / 2}} \]

For binary variables

\[ d = \frac{100 \times (\bar{p}_{\text{treat}} - \bar{p}_{\text{control}})}{\sqrt{\bar{p}_{\text{treat}} (1 - \bar{p}_{\text{treat}}) + \bar{p}_{\text{control}} (1 - \bar{p}_{\text{control}}) / 2}} \]
Illustrative Example
We will use this example to illustrate the application

Health Outcomes with and without Use of Inotropic Therapy in Cardiac Surgery

Results of a Propensity Score–matched Analysis

Dorthe Viemose Nielsen, M.D., Malene Kærslund Hansen, M.B.B.S., Søren Paaske Johnsen, M.D., Ph.D., Mads Hansen, M.D., Karsten Hindsholm, M.D., H.D., Carl-Johan Jakobsen, M.D.

This article has been selected for the Anesthesiology CME Program. Learning objectives and disclosure and ordering information can be found in the CME section at the front of this issue.
What is known and unknown about inotropic therapies

What is known:

- Improves hemodynamic outcomes—ventricular function
- BUT, may be associated with increased morbidity in patients with ischemic heart disease *(increased risk of myocardial oxygen consumption)*

What is unknown:

- Effect on hard “clinical” outcomes
- Effect on mortality
In patients undergoing cardiac surgery, does the use of intra- and post-op use of inotropic therapy compared with non-use improve the mortality and post-op complications over 30 days or one year?
Fig. 2. Graphical representation of absolute standardized differences before and after propensity score matching comparing covariate values. Solid fixed vertical line represents the fixed limit of 10% for absolute standardized difference. CABG = coronary artery bypass grafting; CPB = cardio pulmonary bypass; LVEF = left ventricular ejection fraction; RRT = renal replacement therapy.
Table 1. Patient Characteristics of Cohort

<table>
<thead>
<tr>
<th></th>
<th>Original Cohort</th>
<th>Propensity-matched Cohort</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Inotrope Therapy</td>
<td>Inotrope Therapy</td>
<td>No Inotrope Therapy</td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (yr)</td>
<td>66 (58–73)</td>
<td>70 (62–76)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Females</td>
<td>977 (25.0)</td>
<td>654 (31.2)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>EuroSCORE</td>
<td>4.0 (2–7)</td>
<td>8.0 (6–10)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Patient-related EuroSCORE variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic pulmonary disease</td>
<td>357 (9.1)</td>
<td>285 (13.6)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Extracardiac arteriopathy</td>
<td>347 (8.9)</td>
<td>308 (14.7)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Neurologic dysfunction disease</td>
<td>252 (6.4)</td>
<td>169 (8.1)</td>
<td>0.0234</td>
</tr>
<tr>
<td>Previous cardiac surgery</td>
<td>131 (3.4)</td>
<td>32 (1.5)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Serum creatinine &gt;200 μmol/l</td>
<td>56 (1.4)</td>
<td>138 (6.6)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Active endocarditis</td>
<td>50 (1.3)</td>
<td>93 (4.4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Critical preoperative state</td>
<td>80 (2.0)</td>
<td>267 (12.7)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>EuroSCORE—patient factors</td>
<td>0 (0–1)</td>
<td>0 (0–3)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Other patient-related variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preoperative arrhythmia</td>
<td>311 (8.0)</td>
<td>441 (21.0)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Preoperative RRT</td>
<td>17 (0.4)</td>
<td>39 (1.9)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Cardiac-related EuroSCORE variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unstable angina</td>
<td>297 (7.6)</td>
<td>267 (12.7)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Recent myocardial infarction</td>
<td>716 (18.3)</td>
<td>534 (25.5)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Pulmonary hypertension</td>
<td>104 (2.7)</td>
<td>337 (16.1)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LVEF ≤30%</td>
<td>66 (1.7)</td>
<td>399 (19.0)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>EuroSCORE—cardiac factors</td>
<td>0 (0–0)</td>
<td>0 (0–2)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
Used appropriate statistical analyses

✓ Time-to-event outcome (time-to-death)
  ✓ Log-rank test
  ✓ Survival analysis/Cox-regression stratified by matched pairs
  ✓ Report:
    □ Hazard ratio (95%CI);p

✓ Binary outcome (complications)
  ✓ Conditional logistic regression
  ✓ Report:
    □ OR (95%CI);p
Use of inotropes is significantly associated with increased risk of mortality.

### Table 3. Cumulative Incidence Risk and Hazard Ratios for Death by Treatment Status among Matched Cohort

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>Number of Events</th>
<th>Number at Period Start</th>
<th>Cumulative Incidence Risk, % (95% CI)</th>
<th>HR (95% CI)</th>
<th>Adjusted HR* (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-day mortality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No inotropes</td>
<td>16</td>
<td>1,170</td>
<td>1.37 (0.84–2.22)</td>
<td>1 (reference)</td>
<td>1 (reference)</td>
</tr>
<tr>
<td>Intrope treatment</td>
<td>59</td>
<td>1,170</td>
<td>5.06 (3.94–6.48)</td>
<td>3.69 (2.12–6.41)</td>
<td>3.71 (2.11–6.53)</td>
</tr>
<tr>
<td>One-year mortality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No inotropes</td>
<td>49</td>
<td>1,170</td>
<td>4.19 (3.18–5.50)</td>
<td>1 (reference)</td>
<td>1 (reference)</td>
</tr>
<tr>
<td>Intrope treatment</td>
<td>129</td>
<td>1,170</td>
<td>11.06 (9.39–13.01)</td>
<td>2.51 (1.80–3.50)</td>
<td>2.49 (1.78–3.48)</td>
</tr>
</tbody>
</table>

* Adjusted by anesthetist and surgeon provider group (table 2).

HR = hazard ratio.

**Fig. 3.** Cumulative 1-yr mortality risk by treatment status. Log-rank P value < 0.00001.
Use of inotropes is significantly associated with increased risk of post-op complications:

- Renal replacement therapy
- Myocardial infarction
- Stroke
- Arrhythmia

### Table 4. Odds Ratios of In-hospital Complications by Treatment Status among Matched Cohort

<table>
<thead>
<tr>
<th>In-hospital Complications</th>
<th>Number of Events</th>
<th>OR (95% CI)</th>
<th>Adjusted OR* (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RRT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No inotropes</td>
<td>13</td>
<td>1 (reference)</td>
<td>1 (reference)</td>
</tr>
<tr>
<td>Inotrope treatment</td>
<td>79</td>
<td>7.0 (3.72–13.17)</td>
<td>7.89 (3.80–16.42)</td>
</tr>
<tr>
<td><strong>MI</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No inotropes</td>
<td>45</td>
<td>1 (reference)</td>
<td>1 (reference)</td>
</tr>
<tr>
<td>Inotrope treatment</td>
<td>88</td>
<td>2.02 (1.40–2.93)</td>
<td>2.06 (1.41–3.02)</td>
</tr>
<tr>
<td><strong>Stroke</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No inotropes</td>
<td>19</td>
<td>1 (reference)</td>
<td>1 (reference)</td>
</tr>
<tr>
<td>Inotrope treatment</td>
<td>47</td>
<td>2.61 (1.52–4.50)</td>
<td>2.42 (1.37–4.28)</td>
</tr>
<tr>
<td><strong>Arrhythmia</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No inotropes</td>
<td>387</td>
<td>1 (reference)</td>
<td>1 (reference)</td>
</tr>
<tr>
<td>Inotrope treatment</td>
<td>428</td>
<td>1.17 (0.98–1.38)</td>
<td>1.15 (0.97–1.37)</td>
</tr>
</tbody>
</table>

* Adjusted by anesthetist and surgeon provider group (from table 2).

MI = myocardial infarction; OR = odds ratio; RRT = renal replacement therapy.
Best resources on propensity score methods

Peter Austin
Scientist at ICES
A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003

Peter C. Austin¹,²,³,*,,†

¹Institute for Clinical Evaluative Sciences, Toronto, Ont., Canada
²Department of Public Health Sciences, University of Toronto, Toronto, Ont., Canada
³Department of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ont., Canada
Investigation of statistical properties of propensity score methods

American Journal of Epidemiology Advance Access published August 28, 2010

DOI: 10.1093/aje/kw224

Statistical Criteria for Selecting the Optimal Number of Untreated Subjects Matched to Each Treated Subject When Using Many-to-One Matching on the Propensity Score

Peter C. Austin

Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies

Peter C. Austin

The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies

Peter C. Austin, PhD

Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples

Peter C. Austin

Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples

Peter C. Austin
A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use

Peter C. Austin\textsuperscript{1,2,3,*,†} and Muhammad M. Mamdani\textsuperscript{1,3,4}

A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study

Peter C. Austin\textsuperscript{1,2,3,*,†}, Paul Grootendorst\textsuperscript{4} and Geoffrey M. Anderson\textsuperscript{1,3}

A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality

Peter C. Austin

\textit{Institute for Clinical Evaluative Sciences and University of Toronto}
Key Limitations of the method

- Retrospective design of the study
- Does not work well if
  - Sample size or # of events is small
  - # of confounders is large
- Unreliable when Rx and control groups differ markedly in covariates distributions
Take home Message:

- Propensity score methods can help us to:
  - Correct the effects of channeling bias in observational studies
  - Approximate balance of baseline covariates at the analysis stage

- Careful planning can enhance chances to obtain reliable, accurate and precise estimates of causal relationships

- None of the methods (PS, regression, matching, etc) can reproduce the RCT results
Randomization is the only known method of balancing known and unknown factors between treatment arms in a trial

✓ to ensure comparability
✓ to minimize selection bias
Insert examples where propensity score methods analyses showed different results from those of subsequent RCTs
Show example where propensity score methods analyses showed same results as multiple regression
IMPORTANT

Good observational studies are “designed”
Thank You!