Observational & Quasi-experimental Research Methods

10th Annual Kathleen Foley Palliative Care Retreat
Old Québec, October 24, 2016

Melissa M. Garrido, PhD¹ and Jay Magaziner, PhD²

1. Department of Veterans Affairs, Icahn School of Medicine at Mount Sinai, New York, NY
2. University of Maryland School of Medicine, Baltimore, MD
Overview of Workshop

• Randomized trials and observational studies

• Selection bias – what is it and why do we care?

• Whirlwind tour of some methods to address selection bias
  – Propensity scores
  – Coarsened exact matching
  – Instrumental variables

• Q & A
When our objective is to understand the effect of a treatment or management strategy on an outcome:

Options: Experiment or Observational Study
Considerations: Trade offs between approaches
Experiments

- Best suited for evaluating efficacy
- Can be used to evaluate effectiveness
Randomization

Fundamental Procedure in an Experiment
The Clinical Trial Paradigm

1. Target Population
2. Screened Population
   - Eligibility and Consent
3. Study Population
   - Randomization
4. Treatment Group 1 Patients
5. Treatment Group 2 Patients
   - Treatment and Follow-Up

Outcomes Ascertained

Data Analysis: Comparison of Outcomes
Study Population

Study Population

Randomization

Treatment Group 1 Patients

Treatment Group 2 Patients

Treatment and Follow-Up
Randomization and Assignment

• Randomization is a scientific principle
• Assignment (Allocation) is by chance
• Assures that two groups are alike
Randomization Influence on Clinical Trial

- All aspects
- Definitive for Internal Validity
- Valid Statistical Tests without need to adjust for confounders because groups are alike in all ways (even those that are not easily measured or correlated with those that are easily measured)
Desired Features of Allocation

• Unpredictable
• Avoids selection effects (aka, confounding by indication)
• Secure (Not Switchable)
Intention to Treat Approach

Include anyone who was randomized in the group to which they were randomized.

• Advantages:
  – Provides the most fair comparison between groups because the groups should be equivalent with respect to prognostic factors due to randomization.
  – Straight-forward to implement, no subjective judgment or additional information is required.
Intention to Treat Approach (continued)

• Disadvantages:
  – Results in an estimate of the rate of events among those assigned to a treatment group, whereas, there might be greater interest in the rate of events among those who receive a particular treatment.
  – Includes individuals who did not get treatment.
  – Potential problem if a lot of people change treatment after randomization and before outcome is assessed.
Per-Protocol Analysis

Only include in the analysis those who received the treatment to which they were randomized.

• Advantages:
  – Provides an estimate of the rate of events among those who actually receive the treatment.

• Disadvantages:
  – Could result in a biased estimate of the impact of the treatment in the population because those who do not receive the treatment could be different than those who do.
  – Could result in a biased estimate of treatment differences because those excluded from one group may differ from those excluded from another group.
As-treated Analysis

Include all those in the study in the groups based on treatment received, not treatment randomized to.

• Advantages:
  – Similar to per-protocol but includes more patients.

• Disadvantages:
  – Similar to per-protocol, but is even more likely to result in differences between groups
  – Much like an observational study with a pre-selected group of participants.
Effectiveness vs efficacy

**Effectiveness**
- Generalizable to wider population
- External validity
- Generally less costly
- Can be done when experiment is not an option
- Not practical when treatment or outcome occur rarely in population being studied
- Other

**Efficacy**
- Generalizable only to those meeting entry criteria
- Internal validity
- Generally more costly
- Not always feasible
- Can be done as long as eligible patient group can be identified and enrolled
- Other
Making Sense of Observational Data

Correlation: 99%  Sources: US Census & USDA  tylervigen.com
Confounding due to Selection Bias in Observational Data

Patients not randomized to treatment

Patient characteristics may be associated with both participation in treatment and outcome
True Treatment Effect: 25% improvement in satisfaction with care

Impact of Selection Bias on Analytic Inferences

- Estimate that does not account for selection bias (less likely to capture true treatment effect)
- Estimate that accounts for selection bias (more likely to capture true treatment effect)
Tools to Address Confounding

- Multivariable models
- Matching
- Propensity scores
- Instrumental variables
- Regression discontinuity
- Difference-in-differences
Tools to Address Confounding due to Selection Bias

• **Matching**
  – Compare treated and comparison individuals who have the same values for a set of covariates

• **Propensity scores**
  – Compare treated and comparison individuals who have similar “propensities” or likelihoods for receiving treatment, conditional on a set of several covariates

• **Instrumental variables**
  – Include an additional variable in your model (the “instrument”) that is associated with treatment likelihood but not with outcome
Addressing Selection Bias with Exact Matching

- Goal: Match patients so well that you could imagine that they were randomly assigned to each group

- For each patient in the treatment group, find at least one untreated patient from the comparison group who is identical or as similar as possible on all baseline characteristics

- By matching patients at the individual level, the treatment and comparison groups will be matched at the group level
Matching on Specific Variables:

Match on gender and age

Comparison group
35-44
1 chronic
35-44
2 chronic

Treatment group
35-44
0 chronic
35-44
0 chronic

45-54
1 chronic
45-54
2 chronic

55-64
0 chronic
55-64
1 chronic
55-64
2 chronic

55-64
2 chronic
55-64
0 chronic
55-64
3+ chronic

55-64
3+ chronic

55-64
3+ chronic

55-64
3+ chronic

55-64
3+ chronic

35-44
3+ chronic
35-44
2 chronic

45-54
0 chronic
45-54
2 chronic

45-54
3+ chronic
45-54
3+ chronic

45-54
3+ chronic

45-54
3+ chronic

45-54
3+ chronic

45-54
3+ chronic
Matching on Specific Variables:
Gender, age, number of chronic conditions

Comparison group  
35-44: 0 chronic, 1 chronic, 2 chronic, 3+chronic
45-54: 0 chronic, 1 chronic, 2 chronic, 3+chronic
55-64: 0 chronic, 1 chronic, 2 chronic, 3+chronic
35-44: 0 chronic, 1 chronic, 2 chronic, 3+chronic
45-54: 0 chronic, 1 chronic, 2 chronic, 3+chronic

Treatment group
Isn’t There an Easier Way?

Couldn’t we match on a single composite score instead?

Propensity Score Matching
Propensity Scores: Big Picture

• Create a single composite score of all observed, measured potential confounders of the association between treatment and outcome

• Propensity score is the conditional probability of treatment given the observed covariates X

\[ E(X) = P(D=1 \mid X) \]

• Match or weight on this one-dimensional score alone

• Do this without knowledge of the outcome variable
Propensity Score Assumption: Strongly Ignorable Treatment Assignment

• Given a set of covariates:
  – Treatment assignment and outcome are independent
  – Everyone has a nonzero chance of receiving the treatment

What Propensity Scores Can & Cannot Do

• Propensity scores can:
  – Help find matches from comparison group so that measured confounders are equally distributed between treatment & comparison groups
  – Improve precision of treatment effect estimates

• Propensity scores cannot:
  – Account for unmeasured confounders
General Procedure

Step 1: Choose variables to include in propensity score

Step 2: Ensure that propensity score is balanced across treatment and comparison groups

Step 3: Ensure that covariates are balanced across treatment and comparison groups within blocks of the propensity score

Step 4: Choose a matching or weighting strategy

Step 5: Ensure that covariates are balanced across treatment and comparison groups in sample matched or weighted by propensity score

Step 6: Proceed with analyses based on sample matched or weighted by propensity score

Calculating a propensity score is an iterative process. Steps 1-5 may be repeated several times.

Garrido et al. 2014. HSR 49: 1701-1720
General Procedure

Step 1: Choose variables to include in propensity score

List potential confounders

Step 2: Ensure that propensity score is balanced across treatment and comparison groups

Evaluate feasibility of including these confounders

Step 3: Ensure that covariates are balanced across treatment and comparison groups within blocks of the propensity score

Calculate propensity score with logit or probit regression

Step 4: Choose a matching or weighting strategy

Step 5: Ensure that covariates are balanced across treatment and comparison groups in sample matched or weighted by propensity score

Step 6: Proceed with analyses based on sample matched or weighted by propensity score
Choosing Variables for Propensity Scores

• **Include:**
  – Theoretically related to treatment and outcome
  – Available & easy/reliable to collect on everyone
  – Correlated with unmeasured confounders

• **Do not include:**
  – Variables hypothesized to be associated with treatment but not with outcome
  – Variables that may be affected by the treatment
  – Variables that predict treatment status perfectly
Variable Selection Example

• Hospitalized veterans receiving a palliative care consultation in a VISN 3 acute care facility

• Treatment: Psychotherapy provided after a palliative care consultation

• Outcome: All-cause 30-day readmission
Choosing Variables for Propensity Score Models

1. Pre-Admission
2. Hospital Admission
3. Palliative Care Consultation
4. Mental Health Care Post-PC Consultation (Treatment Group)
5. Discharge
6. No Mental Health Care Post-PC Consultation (Comparison Group)
7. 30 Days Post-Hospital Discharge

Variables occurring before treatment can be included in the propensity score.

Treatment variable

Variables occurring after treatment but before outcome is measured are part of causal pathway and should not be included in the propensity score.

Outcome: Incidence of readmission

Garrido 2014. JPSM 48:711-718
Working Example

- 2008 Healthcare Cost and Utilization Project (HCUP) Nationwide Inpatient Sample
  – Discharge data for hospitalizations throughout the US

- 12,686 patients with metastatic cancer who died during the hospitalization

- Treatment: Palliative Care Consultation

- Outcome: Average total charges per day

- Contrived example – Please do not draw any conclusions from data presented here!
Calculate Propensity Score

- Maximum Likelihood Estimation (logit, probit models)
- Generalized Boosting Methods
- Generalized Method of Moments (Covariate Balancing Propensity Score [CBPS])
Stata Code to Calculate Propensity Score

```
pscore treatment covariate1 covariate2 ... covariate#
pscore(pc_pscore) blockid(pc_block) detail
```

- **Treatment variable**
- **Specify covariates to include in calculation**
- **Label the estimated propensity score**
- **Label the blocks of propensity scores**
- **Optional command that shows details of testing blocks and balancing covariates**

*pscore is not part of Stata’s built-in commands. Type “findit pscore” in Stata’s command line and follow link in pop-up window to install (st0026, Becker & Ichino)*
Working Example: Propensity Score

Treatment variable

Specify covariates to include in calculation

Label the estimated propensity score

Label the blocks of the propensity score

```
pscore = palliative ///
age35to44 age45to54 age55to64 age65to74 age75andup female race_Black ///
race_Hispanic race_other race_missing pay_Medicare pay_Medicaid pay_outofpocket ///
pay_otherormiss NCHS2 NCHS3 NCHS4 NCHS5 NCHS6 ///
lung_ca liver_ca pancreas_ca leukemia stom_ca ///
septicemia pneumonia respfailure renalfailure, ///
pscore(pc_pscore) blockid(pc_block) detail
```
Beginning of output from pscore command

Algorithm to estimate the propensity score

The treatment is palliative

<table>
<thead>
<tr>
<th></th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10,861</td>
<td>86.12</td>
<td>86.12</td>
</tr>
<tr>
<td>1</td>
<td>1,751</td>
<td>13.88</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>12,612</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

Estimation of the propensity score

Iteration 0: log likelihood = -5080.6799
Iteration 1: log likelihood = -4839.3575
Iteration 2: log likelihood = -4836.4592
Iteration 3: log likelihood = -4836.4577

Probit regression

Number of obs = 12612
LR chi2(28) = 488.44
Prob > chi2 = 0.0000
Pseudo R2 = 0.0481

Log likelihood = -4836.4577

| palliative | Coef.     | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|------------|-----------|-----------|-------|-----|---------------------|
| age35to44  | -.1966959 | .1062481  | -1.85 | 0.064 | (-.4049385, .0115466) |
| age45to54  | -.1114731 | .0834791  | -1.34 | 0.182 | (-.2750892, .0521429) |
| age55to64  | -.1496492 | .0780981  | -1.92 | 0.055 | (-.3027187, .0034203) |
| age65to74  | -.0636428 | .077005   | -0.83 | 0.409 | (-.2145698, .0872842) |
| age75andup | -.078232  | .076199   | -1.03 | 0.305 | (-.2275793, .0711153) |
| female     | .055816   | .028897   | 1.93  | 0.053 | (.0008225, .1124545) |
| race_Black | -.1218066 | .0495814  | -2.46 | 0.014 | (-.2189843, -.0246288) |
| race_Hispa-c| -.0708756 | .0667917  | -1.06 | 0.289 | (-.201785, .0600337) |
| race_other | .1128539  | .0595336  | 1.90  | 0.058 | (.0038297, .2295376) |
| race_missing| -.2662183 | .0414828  | -6.42 | 0.000 | (-.3475232, -.1849135) |
| pay_Medicare| -.193896  | .0413946  | -4.68 | 0.000 | (-.2750279, -.1127641) |
| pay_Medicaid| -.0507248 | .056659   | -0.90 | 0.371 | (-.1617744, .0603247) |
| pay_outofp-t| .0283215  | .0911258  | 0.31  | 0.756 | (-.1502818, .2069248) |
| pay_othero-s| .4867789  | .0666438  | 7.30  | 0.000 | (.3561594, .6173984) |
| NCHS2      | -.072554  | .0407098  | -1.78 | 0.075 | (-.1523436, .0072357) |
| NCHS3      | .1675837  | .0417866  | 4.01  | 0.000 | (.0856836, .2494839) |
| NCHS4      | .1728791  | .0562403  | 3.07  | 0.002 | (.0626502, .283108)  |
| NCHS5      | .0499702  | .0509013  | 1.00  | 0.318 | (-.048207, .1481474) |
| NCHS6      | .0169394  | .0596206  | 0.28  | 0.776 | (-.0999148, .1337936) |
Working Example: Propensity Score

The `-pscore-` command provides you with a single score on which to match your treatment and comparison groups.

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
<th>Estimated propensity score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>0.0330385</td>
<td>Obs</td>
</tr>
<tr>
<td>5%</td>
<td>0.0484464</td>
<td>Sum of Wgt.</td>
</tr>
<tr>
<td>10%</td>
<td>0.0607721</td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>0.0876162</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>0.1289967</td>
<td></td>
</tr>
<tr>
<td>75%</td>
<td>0.1775761</td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>0.2239398</td>
<td></td>
</tr>
<tr>
<td>95%</td>
<td>0.2577822</td>
<td></td>
</tr>
<tr>
<td>99%</td>
<td>0.3864837</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Largest</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5015948</td>
<td>0.1388897</td>
<td>0.0695261</td>
</tr>
<tr>
<td>0.50889</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5122628</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5437794</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0048339</td>
<td>1.220613</td>
<td>5.721294</td>
</tr>
</tbody>
</table>
General Procedure

Step 1: Choose variables to include in propensity score

Step 2: Ensure that propensity score is balanced across treatment and comparison groups

Step 3: Ensure that covariates are balanced across treatment and comparison groups within blocks of the propensity score

Step 4: Choose a matching or weighting strategy

Step 5: Ensure that covariates are balanced across treatment and comparison groups in sample matched or weighted by propensity score

Step 6: Proceed with analyses based on sample matched or weighted by propensity score

Check range of common support
Check balance of propensity score
Check Range of Common Support

Extent to which distributions of propensity scores in treatment and comparison groups overlap

`psgraph, treated(treatment) pscore(pc_pscore)`
Check Balance of Propensity Score Across Groups

- Does the propensity score have a similar distribution across treatment and comparison groups?
- Estimate distribution by splitting sample by quintiles or other strata of propensity score
- Test whether mean of propensity score is equal in treatment and comparison groups within each quintile
- If not equal, split one or more quintiles into smaller blocks and compare means
Stata stratifies your data based on the propensity score.

Tests whether mean propensity score is equal for treated and controls within each block.

Groups are significantly different.
Stata Output for Propensity Score Balance
(Continuation of -pscore- output, with “detail” option specified)

Stata splits Block 1 into two blocks and tests whether the propensity score is different for treated and controls in the new Block 1.

Groups are still significantly different.

Stata will automatically continue to split blocks and perform t-tests until it calculates the smallest # of blocks where the propensity score is equivalent across treated and controls in each block.
General Procedure

Step 1: Choose variables to include in propensity score

Step 2: Ensure that propensity score is balanced across treatment and comparison groups

Step 3: Ensure that covariates are balanced across treatment and comparison groups within blocks of the propensity score

Step 4: Choose a matching or weighting strategy

Step 5: Ensure that covariates are balanced across treatment and comparison groups in sample matched or weighted by propensity score

Step 6: Proceed with analyses based on sample matched or weighted by propensity score
Check Balance of Covariates within Blocks of the Propensity Score

- Ideally, for each unique value of the propensity score, the distribution of $X$ (composite of all covariates) is the same for the treatment and comparison groups.

- This is practically impossible, so we check the balance of each observed covariate within blocks of the propensity score.
Stata Output for Propensity Score Balance

(Continuation of -pscore- output, without “detail” option)

---

Step 1: Identification of the optimal number of blocks
use option detail if you want more detailed output

The final number of blocks is 9
This number of blocks ensures that the mean propensity score
is not different for treated and controls in each blocks

Step 2: Test of balancing property of the propensity score
use option detail if you want more detailed output

---

Step 2 is completed (propensity score balanced across groups)

Stata uses t-tests to determine whether each covariate is balanced within each block

You will usually get an error message

---

<table>
<thead>
<tr>
<th>Inferior of block of p-score</th>
<th>palliative 0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>670</td>
<td>27</td>
<td>697</td>
</tr>
<tr>
<td>.05</td>
<td>1,437</td>
<td>104</td>
<td>1,541</td>
</tr>
<tr>
<td>.075</td>
<td>1,633</td>
<td>170</td>
<td>1,803</td>
</tr>
<tr>
<td>.1</td>
<td>3,306</td>
<td>460</td>
<td>3,766</td>
</tr>
<tr>
<td>.15</td>
<td>1,283</td>
<td>222</td>
<td>1,505</td>
</tr>
<tr>
<td>.175</td>
<td>1,058</td>
<td>239</td>
<td>1,297</td>
</tr>
<tr>
<td>.2</td>
<td>1,285</td>
<td>389</td>
<td>1,674</td>
</tr>
<tr>
<td>.3</td>
<td>140</td>
<td>90</td>
<td>230</td>
</tr>
<tr>
<td>.4</td>
<td>49</td>
<td>50</td>
<td>99</td>
</tr>
<tr>
<td>Total</td>
<td>10,861</td>
<td>1,751</td>
<td>12,612</td>
</tr>
</tbody>
</table>
Stata Output for Propensity Score Balance
(Continuation of -pscore- output, with “detail” option)

Testing the balancing property for variable respfailure in block 2
Two-sample t test with equal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1437</td>
<td>0.7974948</td>
<td>0.0160649</td>
<td>0.4020066</td>
<td>0.7766921 - 0.8182975</td>
</tr>
<tr>
<td>1</td>
<td>104</td>
<td>0.6826923</td>
<td>0.04586</td>
<td>0.4676822</td>
<td>0.5917398 - 0.7736448</td>
</tr>
<tr>
<td>combined</td>
<td>1541</td>
<td>0.7897469</td>
<td>0.0103838</td>
<td>0.4076206</td>
<td>0.7693791 - 0.8101147</td>
</tr>
<tr>
<td>diff</td>
<td>0.1148025</td>
<td>0.0413015</td>
<td>0.033793</td>
<td>0.1958157</td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{diff} = \text{mean}(0) - \text{mean}(1) \]
\[ \text{Ho: } \text{diff} = 0 \]
\[ \text{degrees of freedom} = 1539 \]
\[ t = 2.7796 \]

\[ \Pr(|t| > |t|) = 0.0055 \]
\[ \Pr(T > t) = 0.0028 \]

Variable respfailure is not balanced in block 2

Testing the balancing property for variable renalfailure in block 2
Two-sample t test with equal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1437</td>
<td>0.6256089</td>
<td>0.0127714</td>
<td>0.4841338</td>
<td>0.6005564 - 0.6506614</td>
</tr>
<tr>
<td>1</td>
<td>104</td>
<td>0.663615</td>
<td>0.0465393</td>
<td>0.4748137</td>
<td>0.5711221 - 0.7558081</td>
</tr>
<tr>
<td>combined</td>
<td>1541</td>
<td>0.6281635</td>
<td>0.0123155</td>
<td>0.4834519</td>
<td>0.6040066 - 0.6523204</td>
</tr>
<tr>
<td>diff</td>
<td>-0.0378526</td>
<td>0.0490983</td>
<td>0.1341593</td>
<td>0.0584541</td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{diff} = \text{mean}(0) - \text{mean}(1) \]
\[ \text{Ho: } \text{diff} = 0 \]
\[ \text{degrees of freedom} = 1539 \]
\[ t = -0.7710 \]

\[ \Pr(T < t) = 0.2204 \]
\[ \Pr(|T| > |t|) = 0.4409 \]
\[ \Pr(T > t) = 0.7796 \]

Variable renalfailure is balanced in block 2
Improving the Balance of the Propensity Score

• Some imbalance between the groups is usually expected
• Focus on balance of covariates that are more theoretically important

• Consider interactions/correlations between covariates
• Drop 1 or 2 covariates that are less important
• Re-categorize variables
• Include higher order terms or splines of variables
Improving the Balance of the Propensity Score

1. Drop variables created by Stata for initial run of \texttt{pscore} command
   
   \begin{verbatim}
   drop pc_pscore pc_block
   \end{verbatim}

2. Change covariates

3. Re-run \texttt{pscore} command

\begin{verbatim}
pscore palliative //
age35to44 age45to54 age55to64 age65to74 age75andup female race_Black ///
race_Hispanic race_other race_missing pay_Medicare pay_Medicaid pay_outofpocket ///
pay_otherormiss NCHS2 NCHS3 NCHS4 NCHS5_6 ///
lung_ca liver_ca pancreas_ca leukemia stom_ca ///
septicemia pneumonia respfailure renalfailure, ///
pscore(pc_pscore) blockid(pc_block)
\end{verbatim}
You will usually get an error message

1\textsuperscript{st} try: 4 variables unbalanced in 5 blocks

2\textsuperscript{nd} try: 5 variables unbalanced in 4 blocks
You will usually get an error message

Some imbalance between groups is usually expected
Assess Balance with Standardized Differences

• Account for means and variances
• Not sensitive to sample size

Continuous variables

\[ d = \frac{(\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}})}{\sqrt{\frac{s^2_{\text{treatment}} + s^2_{\text{control}}}{2}}} \]

Dichotomous variables

\[ d = \frac{(\hat{p}_{\text{treatment}} - \hat{p}_{\text{control}})}{\sqrt{\frac{\hat{p}_{\text{treatment}}(1 - \hat{p}_{\text{treatment}}) + \hat{p}_{\text{control}}(1 - \hat{p}_{\text{control}})}{2}}} \]

Balanced Propensity Score

![Histogram showing the distribution of propensity scores for control and treatment groups. The x-axis represents the propensity score, ranging from 0 to 0.4, and the y-axis represents frequency. The bars are color-coded: blue for control and red for treatment. The graph illustrates that the distribution is balanced across the propensity score ranges.]
Balance of Covariates: Caution

• Propensity scores only balance measured confounders

• Balance in measured variables does not indicate balance in unmeasured variables

• Unmeasured confounders will bias treatment effect estimates
Balance of Covariates: Caution

• **Do not use** c-statistics, area under the curve, or any other model fit statistics to measure propensity score performance
  – They **do not** measure reduction in confounding
### General Procedure

<table>
<thead>
<tr>
<th>Step 1: Choose variables to include in propensity score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 2: Ensure that propensity score is balanced across treatment and comparison groups</td>
</tr>
<tr>
<td>Step 3: Ensure that covariates are balanced across treatment and comparison groups within blocks of the propensity score</td>
</tr>
</tbody>
</table>

**Step 4: Choose a matching or weighting strategy**

<table>
<thead>
<tr>
<th>Step 5: Ensure that covariates are balanced across treatment and comparison groups in sample matched or weighted by propensity score</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Step 6: Proceed with analyses based on sample matched or weighted by propensity score</th>
</tr>
</thead>
</table>

Matching and Weighting Strategies

Quality  <->  Quantity

Nearest Neighbor
Radius Matching
Kernel Weighting
Inverse Probability of Treatment Weighting

No universal “best” strategy
Choices When Matching Sample by Propensity Score

• How close of a match is acceptable?

• Should every treated individual have one or many matches in the comparison group?

• Should treated individuals be matched with or without replacement?

• Should matching be greedy or optimal?
Which Strategy to Choose?

• No best method

• **Without examining outcome**, evaluate covariate balance in several strategies (our next step – Step 5)

• Choose the method that has the best balance and still meets the analytic goal
Stata Code to Match Sample on Propensity Score

```stata
qui psmatch2 treatment, outcome(outcomevar)
   pscore(pc_pscore) caliper(.013828) neighbor(1)
```

- **Treatment variable**
- **Dependent variable**
- **Calculated propensity score**
- **Option for caliper matching**
- **Option for number of matches**
Stata Code to Weight Sample on Propensity Score

Kernel Weight:

```stata
qui psmatch2 treatment, kernel outcome(outcomevar) pscore(pc_pscore)
```

IPTW:

```stata
qui dr outcomevar treatment covariatel... covariate#, genvars
```

Creates variable “iptwt” that stores the weights calculated by this command
General Procedure

Step 1: Choose variables to include in propensity score

Step 2: Ensure that propensity score is balanced across treatment and comparison groups

Step 3: Ensure that covariates are balanced across treatment and comparison groups within blocks of the propensity score

Step 4: Choose a matching or weighting strategy

Step 5: Ensure that covariates are balanced across treatment and comparison groups in sample matched or weighted by propensity score

Step 6: Proceed with analyses based on sample matched or weighted by propensity score

Perform multiple checks
Several Ways to Evaluate Balance in Sample Matched or Weighted by Propensity Score

• Standardized differences
• Graphs
  – Quantile-quantile plots
  – Plots of covariates in treated and comparison groups
• Ratios of variance
Evaluate standardized differences in matched sample

Immediately following \( \texttt{psmatch2} \), run:

\[
\texttt{pptest covariate1..covariate#, treated(treatment) both}
\]

This will show balance before and after match (default is to only show after)

![Table showing standardized differences](image)
### Summary of covariate imbalance

#### Before Matching

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
<th>Largest</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>0.4647914</td>
<td>0.4647914</td>
<td>10.00157</td>
<td>9.429134</td>
</tr>
<tr>
<td>5%</td>
<td>1.525013</td>
<td>1.525013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>1.76495</td>
<td>1.76495</td>
<td>obs 23</td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>2.446086</td>
<td>2.21497</td>
<td>sum of wgt. 23</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>5.245142</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75%</td>
<td>17.56547</td>
<td>24.66997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>25.51863</td>
<td>25.51863</td>
<td>variance 88.90857</td>
<td></td>
</tr>
<tr>
<td>95%</td>
<td>25.5416</td>
<td>25.5416</td>
<td>skewness 0.9390208</td>
<td></td>
</tr>
<tr>
<td>99%</td>
<td>31.01482</td>
<td>31.01482</td>
<td>kurtosis 2.455434</td>
<td></td>
</tr>
</tbody>
</table>

#### After Matching

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
<th>Largest</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>0</td>
<td>0</td>
<td>1.575237</td>
<td>1.141444</td>
</tr>
<tr>
<td>5%</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>0.22809</td>
<td>0.22809</td>
<td>obs 23</td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>0.6399435</td>
<td>0.370463</td>
<td>sum of wgt. 23</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>1.423716</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75%</td>
<td>2.5283</td>
<td>2.865676</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>3.245421</td>
<td>3.245421</td>
<td>variance 1.302894</td>
<td></td>
</tr>
<tr>
<td>95%</td>
<td>3.478944</td>
<td>3.478944</td>
<td>skewness 0.360008</td>
<td></td>
</tr>
<tr>
<td>99%</td>
<td>3.72565</td>
<td>3.72565</td>
<td>kurtosis 1.953917</td>
<td></td>
</tr>
</tbody>
</table>

### Summary of mean and median bias before and after matching

<table>
<thead>
<tr>
<th>Sample</th>
<th>Pseudo R2</th>
<th>LR chi2</th>
<th>p&gt;chi2</th>
<th>MeanBias</th>
<th>MedBias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>0.048</td>
<td>482.73</td>
<td>0.000</td>
<td>10.0</td>
<td>5.2</td>
</tr>
<tr>
<td>Matched</td>
<td>0.002</td>
<td>7.12</td>
<td>0.999</td>
<td>1.6</td>
<td>1.4</td>
</tr>
</tbody>
</table>
Visual inspection of standardized differences

`pptest covariatel..covariate#, treated(treatment) both hist`

Optional command to get histogram of covariate balance
Visual inspection of standardized differences

\texttt{pptest covariate1..covariate\#, treated(treatment) both graph}

Optional command to get dot graph of covariate balance
Evaluate standardized differences in weighted sample

- **Kernel**: Uses `-psmatch2-` so can use same procedure as for matched samples

- **IPTW**:

  After running `-dr-` and normalizing weight variable, run:

  \[ \texttt{pbalchk treatment covariate1... covariate#}, \]

  \[ \texttt{wt(norm_weights)} \]

  Name of weight variable created earlier
Output from `-pbalchk`-

<table>
<thead>
<tr>
<th></th>
<th>Mean in treated</th>
<th>Mean in Untreated</th>
<th>Standardised diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>67.47</td>
<td>67.52</td>
<td>-0.003</td>
</tr>
<tr>
<td>female</td>
<td>0.48</td>
<td>0.47</td>
<td>0.008</td>
</tr>
<tr>
<td>race_Black</td>
<td>0.11</td>
<td>0.11</td>
<td>0.011</td>
</tr>
<tr>
<td>race_Hispanic</td>
<td>0.05</td>
<td>0.05</td>
<td>-0.006</td>
</tr>
<tr>
<td>race_other</td>
<td>0.06</td>
<td>0.06</td>
<td>-0.002</td>
</tr>
<tr>
<td>race_missing</td>
<td>0.17</td>
<td>0.18</td>
<td>-0.007</td>
</tr>
<tr>
<td>pay_Medicare</td>
<td>0.58</td>
<td>0.58</td>
<td>0.001</td>
</tr>
<tr>
<td>pay_Medicaid</td>
<td>0.09</td>
<td>0.09</td>
<td>0.005</td>
</tr>
<tr>
<td>pay_outcofpt</td>
<td>0.03</td>
<td>0.02</td>
<td>0.014</td>
</tr>
<tr>
<td>pay_otherc</td>
<td>0.04</td>
<td>0.04</td>
<td>-0.009</td>
</tr>
<tr>
<td>NCHS2</td>
<td>0.21</td>
<td>0.23</td>
<td>-0.033</td>
</tr>
<tr>
<td>NCHS3</td>
<td>0.17</td>
<td>0.17</td>
<td>-0.006</td>
</tr>
<tr>
<td>NCHS4</td>
<td>0.08</td>
<td>0.08</td>
<td>0.021</td>
</tr>
<tr>
<td>NCHS5_6</td>
<td>0.20</td>
<td>0.18</td>
<td>0.047</td>
</tr>
<tr>
<td>lung_ca</td>
<td>0.30</td>
<td>0.29</td>
<td>0.016</td>
</tr>
<tr>
<td>liver_ca</td>
<td>0.02</td>
<td>0.02</td>
<td>0.020</td>
</tr>
<tr>
<td>pancreas_ca</td>
<td>0.05</td>
<td>0.05</td>
<td>0.001</td>
</tr>
<tr>
<td>leukemia_ca</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.009</td>
</tr>
<tr>
<td>stom_ca</td>
<td>0.03</td>
<td>0.02</td>
<td>0.025</td>
</tr>
<tr>
<td>septicemia</td>
<td>0.34</td>
<td>0.34</td>
<td>-0.015</td>
</tr>
<tr>
<td>pneumonia</td>
<td>0.34</td>
<td>0.32</td>
<td>0.023</td>
</tr>
<tr>
<td>respfailure</td>
<td>0.45</td>
<td>0.46</td>
<td>-0.013</td>
</tr>
<tr>
<td>renalfailure</td>
<td>0.32</td>
<td>0.32</td>
<td>-0.009</td>
</tr>
</tbody>
</table>

Same information as %bias in `-pstest-` output, but not expressed as a percentage.
Plots of Covariates in Treated and Comparison Groups

• Plot density of weighted continuous covariate in treated group against density in comparison group

• Subjective comparison

twoway kdensity covariate if treatment [aweight= norm_weights]
|| kdensity covariate if !treatment [aweight= norm_weights]
Step 1: Choose variables to include in propensity score

Step 2: Ensure that propensity score is balanced across treatment and comparison groups

Step 3: Ensure that covariates are balanced across treatment and comparison groups within blocks of the propensity score

Step 4: Choose a matching or weighting strategy

Step 5: Ensure that covariates are balanced across treatment and comparison groups in sample matched or weighted by propensity score

Step 6: Proceed with analyses based on sample matched or weighted by propensity score

Alternative:
Covariate Balancing Propensity Score Method
Covariate Balancing Propensity Scores

• Generalized method of moments to estimate a propensity score model that optimizes covariate balance across treatment groups
• Typically used with IPTW

• Advantage: Less subject to investigator bias
• Disadvantage: No control over relative weight provided to confounders

• Software – R package ‘CBPS’

General Procedure

Step 1: Choose variables to include in propensity score

Step 2: Ensure that propensity score is balanced across treatment and comparison groups

Step 3: Ensure that covariates are balanced across treatment and comparison groups within blocks of the propensity score

Step 4: Choose a matching or weighting strategy

Step 5: Ensure that covariates are balanced across treatment and comparison groups in sample matched or weighted by propensity score

Step 6: Proceed with analyses based on sample matched or weighted by propensity score
Analysis of Data Matched or Weighted by Propensity Score

- Delete observations from individuals not within the range of common support
- Choose the treatment effect of interest
- Calculate correct standard error for propensity score matched or weighted sample
- Guard against misspecification of the propensity score
Delete Individuals Outside of the Range of Common Support

<table>
<thead>
<tr>
<th>psmatch2: Treatment assignment</th>
<th>psmatch2: Common support on support</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Untreated</td>
<td>10,070</td>
<td>10,070</td>
</tr>
<tr>
<td>Treated</td>
<td>1,574</td>
<td>1,574</td>
</tr>
<tr>
<td>Total</td>
<td>11,644</td>
<td>11,644</td>
</tr>
</tbody>
</table>

No unmatched individuals

<table>
<thead>
<tr>
<th>psmatch2: Treatment assignment</th>
<th>psmatch2: Common support off support</th>
<th>psmatch2: Common support on support</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Untreated</td>
<td>0</td>
<td>10,070</td>
<td>10,070</td>
</tr>
<tr>
<td>Treated</td>
<td>106</td>
<td>1,468</td>
<td>1,574</td>
</tr>
<tr>
<td>Total</td>
<td>106</td>
<td>11,538</td>
<td>11,644</td>
</tr>
</tbody>
</table>

106 treated individuals will be deleted
Treatment Effects

• ATT: Average Treatment Effect on the Treated

• ATE: Average Treatment Effect for sample within range of common support
  - Incorporates ATT and average treatment effect on untreated
Need to Correct Standard Errors for Treatment Effect Estimates

• Uncertainty from propensity score estimate influences uncertainty of treatment effect estimate

• Ignoring uncertainty
  – Makes standard errors for ATEs more conservative (might conclude that there was no evidence of a significant treatment effect when there was)
  – Can make standard errors for ATTs more conservative or more generous
How to Correct Standard Errors

• Do nothing
  – If propensity score and treatment effect are estimated simultaneously, no need for further correction

• Bootstrap
  – When propensity score created in a separate step from treatment effect estimate and sample is weighted by propensity score

• Abadie-Imbens method
  – When propensity score created in a separate step from treatment effect estimate and sample is matched by propensity score
Guarding Against Misspecification of the Propensity Score

- "Doubly-robust" estimation

  - Perform multivariable regression analysis on a sample matched or weighted by the propensity score (Stata’s `aweight` and `pweight` commands)

  - As long as either the propensity score or the regression model is specified correctly, the treatment effect estimates will not be biased
Interpreting Propensity Score Analysis Results

• Generalizability
  – Excluded individuals differ from those within the range of common support
  – Treated and comparison individuals with missing values for any variables used in the propensity score are usually deleted

• Meaning of other coefficients in the model
  – Would need to create a new propensity score to test other interventions in the dataset
Sensitivity Analyses for Residual (Unobserved) Confounding

- Identify smallest amount of unobserved confounding that would need to exist to change your inference from rejection to acceptance of $H_0$

- Test effect of treatment variable on a lagged outcome

- Estimate treatment effect in multiple comparison groups

Liu et al 2013. Prevention Science 14: 570-580
“Pre-processing” Datasets

Make treatment and comparison group as similar as possible on observed confounders before proceeding with analysis

• Propensity Scores
• Exact Matching
• Coarsened Exact Matching
• Entropy Matching

Ho et al. 2007. Political Analysis 15: 199-236
Coarsened Exact Matching

• Match on broad categories (coarsened values) of important variables

• More feasible than exact matching on large set of potential confounders

• Not susceptible to worsened balance due to model misspecification (a strong risk with propensity score matching when data on important confounders are not available)

**Exact Matching**
Match on exact age and number of chronic conditions

- **Comparison group**
  - 50, 0 chronic
  - 54, 1 chronic
  - 55, 2 chronic
  - 56, 2 chronic

- **Treatment group**
  - 51, 0 chronic
  - 54, 1 chronic
  - 55, 2 chronic
  - 56, 2 chronic

**Coarsened Exact Matching**
Match on age category (50-59, 60-69) and presence of chronic conditions

- **Comparison group**
  - 50, 0 chronic
  - 54, 1 chronic
  - 55, 2 chronic
  - 56, 2 chronic

- **Treatment group**
  - 51, 0 chronic
  - 54, 1 chronic
  - 55, 2 chronic
  - 56, 2 chronic
Coarsened Exact Matching Procedure

• Divide sample into strata that have treated and comparison individuals with the same coarsened values of covariates

• Within strata,
  – Treated individuals assigned a weight of 1
  – Comparison individuals are assigned a weight that accounts for the number of: treated observations within the strata, comparison observations within the strata, matched treated observations within the dataset, and matched comparison observations within the dataset

• Strata without both treated and comparison individuals are assigned a weight of 0

• Traditional multivariable analyses are run on the weighted dataset
What CEM Can & Cannot Do

• CEM can:
  – Help find matches from comparison group so that *measured* confounders can be equally distributed between treatment & comparison groups
  – Improve precision of treatment effect estimates

• CEM cannot:
  – Account for *unmeasured* confounders
Stata Code to Perform CEM

**Specify covariates to include in calculation**

cem covariate1 (cutpoint1 cutpoint2..cutpoint3) covariate2 ... covariate#, **treatment**(treatmentvar)

**Specify coarsened values of continuous variables**

*–cem– is not part of Stata’s built-in commands. Type “findit cem” in Stata’s command line and follow link in pop-up window to install.*
Working Example: CEM

Specify covariates to include in calculation:
- Coarsen age into ≤ 65 versus > 65
- Coarsen number of chronic diseases into 0, 1-2, and 3+

Treatment variable:
- Treatment (palliative)
Working Example: Stata Output for CEM

```
.cem age (65.5) female number chronic (0.5 2.5), treatment (palliative)
(using the scott break method for imbalance)

Matching Summary:  
-------------------
Number of strata: 9  
Number of matched strata: 8

0  1  
All 10897 1789  
Matched 10895 1789  
Unmatched 2 0

2 unmatched observations (weight = 0)

Multivariate L1 distance: 0.26375911

Univariate imbalance:

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>mean</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.05185</td>
<td>-0.05637</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>female</td>
<td>2.3e-14</td>
<td>3.9e-14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>number chronic</td>
<td>0.0817</td>
<td>-0.45616</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
</tr>
</tbody>
</table>

0 = perfect balance, 1 = complete imbalance  
**Interpret relative to output from other matches**
```
As variables become less coarsened, finding matches for every observation becomes more difficult.

Multivariate L1 distance: .43637651
CEM: Run Planned Analyses on Weighted Sample

- `cem` produces variable `cem_weights`

- Use un-coarsened values of variables used for matching

Continuous variables can be returned to original form

```
.glm TOTCHG i.palliative age female i.hospsize_small i.hospsize_med i.hosp_private_nfp ///
> i.hosp_private_fp numberchronic [pweight = cem_weights], family(gamma) link(log)
```

Weight command
Interpreting Results of Analyses Using CEM

• Generalize to individuals similar to those included in the matched sample

• ATT
Tools to Address Confounding

- Multivariable models
- Matching
- Propensity scores
- **Instrumental variables**
- Regression discontinuity
- Difference-in-differences
Instrumental Variable Analyses

Instrumental variable (IV): Day of week of hospital admission

Illness severity

Palliative Care

Quality of Life Hospital Readmission Rates

“...Finding a little RCT inside a lot of observational data”

Pizer 2016. HSR. 51: 790-811
What Makes a Good Instrument?

• Related to treatment likelihood
  – *F*-statistic and partial $r^2$

• Not independently related to outcome (exclusion restriction)
  – *Falsification tests*

• Unrelated to other patient characteristics
  – *Standardized differences*

What IV Analysis Can & Cannot Do

• IV analysis can:
  – Reduce selection bias due to both measured and unmeasured confounders
  – Estimate treatment effect for individuals who may or may not get treatment, depending on the value of the IV

• IV analysis cannot:
  – Generalize to individuals who would not be sensitive to the value of the instrumental variable
Instrumental Variable Methods: Two-Stage Least Squares (2SLS) and Control Functions

• Step 1: Model treatment likelihood, include instrumental variable

• Step 2: Model outcome
  – 2SLS: Include treatment likelihood from Step 1
  – Control Function: Include a function of the residuals from Step 1
Stata Code to Perform 2SLS

```
ivreg2 outcomevar (treatmentvar = IV) covariate1 ... covariate#, ffirst
```

Specify this option to get statistics on IV performance without the rest of the first-stage equation output.

*`-ivreg2-` is not part of Stata’s built-in commands. Type “`ssc install ivreg2, replace`” in Stata’s command line*
Estimates from Control Functions More Efficient than Estimates from Two-Stage Least Squares Models

Control Function Estimate

2SLS Estimate

True Treatment Effect: -$100

0
Stata Code for Control Functions: Two-stage residual inclusion

Model treatment likelihood, include IV

```
quiglm ivreg2 treatmentvar IV covariatem1 ... covariatem#, f(family) link(link)
```

Predict the residual from the treatment likelihood equation

```
predict treatment_res, response
```

Model outcome, include residual from treatment likelihood equation

```
glm outcomevar treatmentvar covariatem1..covariatem# treatment_res, f(family) link(link)
```

Generate marginal effects and calculate bootstrapped standard errors
Interpreting Results of IV Analyses

• Generalize to individuals similar to those included in the matched sample

• Local ATE or local ATT
Falsification tests

• Cannot prove the exclusion restriction

• Falsification tests can strengthen argument that exclusion restriction is valid

• Rerun analyses in situations where treatment should not have an effect, but potential confounders might have an effect
  – Alternate outcome
  – Alternate population

• If no evidence of an effect from confounders, strengthens confidence in IV results
Summary

• Observational data can be rich source of information for improving patient outcomes

• Many tools to improve treatment effect estimation from observational data

• Important to understand assumptions, generalizability, and limitations of each tool
Questions?
Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine 2009;28: 3083-3107.


Pizer S. Falsification testing of instrumental variables methods. HSR 2016; 51(2): 790-811.


Acknowledgements

• VA HSR&D (CDA 11-201, CDP 12-255)
• PC4C Study Group (NCI/NINR 5R01CA116227, National Palliative Care Research Center)
• Melissa Aldridge, PhD MBA; Helene Starks, PhD
• Icahn School of Medicine at Mount Sinai Claude D. Pepper Older Americans Independence Center (NIH/NIA P30 AG028741-01A2)

The views expressed in this presentation are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.