For *Objective* Causal Inference, Design Trumps Analysis

Donald B. Rubin
Harvard University
Clear separation between:

- “Science” (and ...)

<table>
<thead>
<tr>
<th>X</th>
<th>Y(0)</th>
<th>Y(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Units

<table>
<thead>
<tr>
<th>N</th>
<th></th>
<th></th>
</tr>
</thead>
</table>

X = Covariates unaffected by treatments
Y(0) = Potential outcomes under control treatment
Y(1) = Potential outcomes under active treatment

Notation due to Neyman (1923) in context of randomized experiments
Classical Experimental Design

• Clear separation between Science and what we do to learn about Science:
  – Randomized assignment of treatments
  – $W = \text{Vector of } N \text{ treatment indicators}$

• This distinction and the consequential clarity should be maintained and should not be forgotten when designing observational studies

• Cochran & Stat 140, great advantage to start here
Causal Inference is a Missing Data Problem

RCM (Holland, 1986) for work in 1970’s
Maintains critical distinction from experimental design

• Same notation for science whether try to learn about it from randomized experiment or observational study.

• Earlier, “$Y_{obs}$” used in nonrandomized studies with $W$ a predictor in regressions, paths, arrows

$$Y_{obs} = \{Y_{obs,i}\}$$

$$Y_{obs,i} = W_i Y_i(1) + (1-W_i)Y_i(0)$$

Entangles Science and assignments
Assignment Mechanism

Creates missing potential outcomes

\[ \Pr(W|X,Y(0),Y(1)) \quad (AM) \]

Randomized experiments are:

- unconfounded: AM = \Pr(W|X), and
- probabilistic: 1 > \Pr(W_i|X_i) > 0 for all i

Earlier, words describing AM but no explicit mathematical notation or expressions (e.g., Roy, 1953)

Same is true for potential outcomes before Neyman’s (1923) notation (e.g., Fisher, 1918)
Design Observational Studies to Approximate Randomized Trials

1. Hide outcome data until the design phase is complete
2. Think very carefully about decision makers and the key covariates that were used to make treatment decisions
3. If key covariates are not observed or very noisy, usually best to give up and seek better data source
4. Find subgroups (subclasses or matched pairs) in which the treatment and control groups have balance – essentially the same distribution of observed covariates
   - Not always possible to achieve balance
   - Inferences are limited to subgroups where balance is achieved

• #1 - #4 combine to create an objective design that approximates a randomized trial in each subclass that is balanced with respect to observed covariates
Cochran (1968) – Illustrative Example with One Key Covariate

- Population: Male smokers in U.S.
- Treatment = cigar/pipe smoking
- Control = cigarette smoking
- Outcome = death rate/1000 person years
- Decision maker is the individual male smoker
- Reason for a smoking male to choose cigarettes versus cigar/pipe?
- **Age** is a key covariate for selection of smoking type for males
Subclassification to Balance Age

• To achieve balance on age, compare:
  – “young” cigar/pipe smokers with “young” cigarette smokers
  – “old” cigar/pipe smokers with “old” cigarette smokers
• Or better, compare:
  – Young, middle aged, old
  – Even more age subclasses
• Design phase, no outcome data, objective:
  – Approximates a randomized trial within subclasses
• Now look at outcome data
### Comparison of Mortality Rates for Two Smoking Groups in U.S.

<table>
<thead>
<tr>
<th></th>
<th>Cigarette Smokers</th>
<th>Cigar/Pipe Smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality Rates per 1000 person-years, %</td>
<td>13.5</td>
<td>17.4</td>
</tr>
<tr>
<td>Adjusted Mortality Rates using subclasses, %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 age subclasses</td>
<td>16.4</td>
<td>14.9</td>
</tr>
<tr>
<td>3 age subclasses</td>
<td>17.7</td>
<td>14.2</td>
</tr>
<tr>
<td>9-11 age subclasses</td>
<td>21.2</td>
<td>13.7</td>
</tr>
</tbody>
</table>


But 20 four-class covariates ⇒ over million million subclasses.
Propensity Score Methods

• Observational study analogue of complete randomization
• The propensity score is the probability of treatment versus control as a function of observed covariates
  – Model the reasons for treatment versus control at the level of the decision makers
  – For example, logistic regression model to predict cigarette versus cigar/pipe smoking with age, education, income, etc. as predictors
• Then subclassify (or match) on the propensity score as if it were the only covariate, e.g., 5-10 subclasses
• If correctly done, this creates balance within each subclass on ALL covariates used to estimate the propensity score
Example: GAO Study of Breast Conservation versus Mastectomy

- Six large and expensive randomized clinical trials had been completed showing little difference for the type of women randomized in the trials and participating clinics
- Question: Same results in U.S. general practice?
- Observational data available
  - SEER Database: covariates, treatments, post-surgery outcomes
- Design phase
  - Hide outcomes
  - Think hard about decision rules and key covariates
  - Key covariates for decisions by doctors/women: Age, marital status, region of country, urbanization, race, size of tumor, etc., all available in SEER and considered sufficient
  - Balance covariates between treatment and control using subclasses
## Estimated 5-year Survival Rates for Node-negative Patients in Six Randomized Clinical Trials

<table>
<thead>
<tr>
<th>Study</th>
<th>Women</th>
<th>Estimated Survival Rate for Women</th>
<th>Estimated Causal Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BC</td>
<td>Mas</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>US-NCI†</td>
<td>74</td>
<td>93.9</td>
<td>94.7</td>
</tr>
<tr>
<td>Milanese†</td>
<td>257</td>
<td>93.5</td>
<td>93.0</td>
</tr>
<tr>
<td>French†</td>
<td>59</td>
<td>94.9</td>
<td>96.2</td>
</tr>
<tr>
<td>Danish‡</td>
<td>289</td>
<td>87.4</td>
<td>85.9</td>
</tr>
<tr>
<td>EORTC‡</td>
<td>238</td>
<td>89.0</td>
<td>90.0</td>
</tr>
<tr>
<td>US-NSABP‡</td>
<td>330</td>
<td>89.0</td>
<td>88.0</td>
</tr>
</tbody>
</table>

†Single-center trial; ‡ Multicenter trial

Estimated 5-year Survival Rates for Node-Negative Patients in the SEER Database within Each of Five Propensity Score Subclasses

<table>
<thead>
<tr>
<th>Propensity Score Subclass</th>
<th>Women</th>
<th>Estimated Survival Rate for Women</th>
<th>Estimated Causal Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Breast Conservation (BC)</td>
<td>Mastectomy (Mas)</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>56</td>
<td>1008</td>
<td>85.6</td>
</tr>
<tr>
<td>2</td>
<td>106</td>
<td>964</td>
<td>82.8</td>
</tr>
<tr>
<td>3</td>
<td>193</td>
<td>866</td>
<td>85.2</td>
</tr>
<tr>
<td>4</td>
<td>289</td>
<td>978</td>
<td>88.7</td>
</tr>
<tr>
<td>5</td>
<td>462</td>
<td>604</td>
<td>89.0</td>
</tr>
<tr>
<td>Averages Across Five Subclasses</td>
<td></td>
<td>86.3</td>
<td>86.9</td>
</tr>
</tbody>
</table>

Diagnostics for Accessing Balance

• Assessing balance simpler in large samples, just as with randomized experiments
• To illustrate diagnostics, use a marketing application that involved a weight loss drug
• Units = doctors
• Treatment = sales rep “visits” doctor to discuss
• Control = no visit
• Decision-makers = sales reps
• Key covariates = prior Rxs, medical specialty, years in practice, size of practice, etc.
Histograms for background variable: Prior Rx Score (0-100) at Baseline

Histograms for background variable: Specialty

Histograms for summarized background variables: Linear Propensity Score

Histograms for a variable in a subclass of propensity scores: Prior Rx Score

Histograms for a variable in a subclass of propensity scores: Specialty

Marketing Example: Achieved Balance

• Within each narrow subclass of propensity scores, the treatment and control groups will be as balanced as if randomly divided

• Claim: This holds for all subclasses in which there are both treated and control subjects, and holds for all covariates that were used to estimate the propensity score

• Works best when the propensity score subclasses have large sample sizes and are relatively narrow

• Five to ten propensity score subclasses often fully adequate to balance all covariates

• No outcome data used in the design stage
Simple Noncompliance, Instrumental Variables, and Bayesian Generalizations

- Template for other observational studies involves more complex randomized experiment
- Illustrate with completely randomized experiment with noncompliance with assigned treatment
- Return later to combined analysis with observational study design
## Sommer and Zeger Vitamin A Data

<table>
<thead>
<tr>
<th>Row</th>
<th>True Compliance Type</th>
<th>Treatment Assignment</th>
<th>Treatment Received</th>
<th>(Y_{obs})</th>
<th>Number of Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>?</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11514</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>74</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2385</td>
</tr>
<tr>
<td>4</td>
<td>N</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>9663</td>
</tr>
<tr>
<td>6</td>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23682</td>
</tr>
</tbody>
</table>

Results of Three Standard MoM Analyses

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
<th>Calculation</th>
<th>Row Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITT</td>
<td>-0.0026</td>
<td>$\frac{12 + 34}{9663 + 2385 + 12 + 34} - \frac{74}{11514 + 74}$</td>
<td>3, 4, 5, &amp; 6 vs. 1 &amp; 2</td>
</tr>
<tr>
<td>As-treated</td>
<td>-0.0065</td>
<td>$\frac{12}{9663 + 12} - \frac{34 + 74}{11514 + 2385 + 34 + 74}$</td>
<td>5 &amp; 6 vs. 1, 2, 3, &amp; 4</td>
</tr>
<tr>
<td>Per protocol</td>
<td>-0.0052</td>
<td>$\frac{12}{9663 + 12} - \frac{74}{11514 + 74}$</td>
<td>5 &amp; 6 vs. 1 &amp; 2</td>
</tr>
</tbody>
</table>

MoM CACE Analysis

\[ \text{ACE} = p_N \cdot \text{NACE} + p_C \cdot \text{CACE} \]

\[-0.0025 = 0.2 \cdot \text{NACE} + 0.8 \cdot \text{CACE} \]

\[-0.0025 = 0.8 \cdot \text{CACE} \quad \Rightarrow \quad \text{CACE} = -0.0025/0.8 = -0.0031 \]
Bayesian Analysis of Sommer & Zeger Data

Bayesian Analysis of Sommer & Zeger Data, Marginal Posterior Distributions with and without Exclusion Restriction

<table>
<thead>
<tr>
<th>Estimand</th>
<th>Exclusion restriction</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Median</th>
<th>5th percentile</th>
<th>95th percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>CACE</td>
<td>No</td>
<td>3.1</td>
<td>2.5</td>
<td>3.2</td>
<td>-0.9</td>
<td>7.0</td>
</tr>
<tr>
<td>ITT$_{(n)}$</td>
<td>No</td>
<td>0.5</td>
<td>10.1</td>
<td>0.2</td>
<td>-14.1</td>
<td>17.5</td>
</tr>
<tr>
<td>CACE</td>
<td>Yes</td>
<td>3.1</td>
<td>1.2</td>
<td>3.1</td>
<td>1.2</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Bayesian Analysis of Sommer & Zeger Data

Bayesian Analysis of Sommer & Zeger Data

Bayesian Analysis of Sommer & Zeger Data

Hypothetical Example Illustrating Frequentist Superiority of Bayes over IVE (MoM) and MLE, Population Parameters with Exclusion Restrictions and Monotonicity

<table>
<thead>
<tr>
<th>$T$</th>
<th>$P(C_i = t \mid \pi)$</th>
<th>$D_i(0)$</th>
<th>$D_i(1)$</th>
<th>$Y_i \mid C_i = t, Z_i = 0, \pi$</th>
<th>$Y_i \mid C_i = t, Z_i = 0, \pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$</td>
<td>0.25</td>
<td>0</td>
<td>1</td>
<td>$N(0.1, 0.16)$</td>
<td>$N(0.9, 0.49)$</td>
</tr>
<tr>
<td>$n$</td>
<td>0.45</td>
<td>0</td>
<td>0</td>
<td>$N(1.0, 0.25)$</td>
<td>$N(1.0, 0.25)$</td>
</tr>
<tr>
<td>$a$</td>
<td>0.30</td>
<td>1</td>
<td>1</td>
<td>$N(0.0, 0.36)$</td>
<td>$N(0.0, 0.36)$</td>
</tr>
</tbody>
</table>

Hypothetical Example Illustrating Frequentist Superiority of Bayes over IVE (MoM) and MLE, One Sample

### Hypothetical Example Illustrating Frequentist Superiority of Bayes over IVE (MoM) and MLE, Frequentist Evaluation under Monotonicity and Exclusion Restrictions

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Mean bias</th>
<th>Median bias</th>
<th>Root mean squared error</th>
<th>Median absolute error</th>
<th>90% interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coverage rate</td>
<td>Median width</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posterior mean</td>
<td>-0.10</td>
<td>-0.07</td>
<td>0.48</td>
<td>0.30</td>
<td>0.91</td>
</tr>
<tr>
<td>Posterior median</td>
<td>-0.08</td>
<td>-0.06</td>
<td>0.51</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>MLE</td>
<td>-0.14</td>
<td>-0.12</td>
<td>0.51</td>
<td>0.31</td>
<td>0.74</td>
</tr>
<tr>
<td>IVE</td>
<td>0.55</td>
<td>0.13</td>
<td>2.31</td>
<td>0.54</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Using a More Complex Template: Randomized Block Experiment with Noncompliance

- Causal effect of Large versus Small treating hospitals on cardia cancer survival
- Dataset from Karolinska Institute, Stockholm
- Medical researchers accept unconfounded assignment of “home (diagnosing) hospital” type, but NOT treating hospital type because of self-selected transfers
- Consider transfers between hospital types as a form of noncompliance with assignment
Using a More Complex Template: Randomized Block Experiment with Noncompliance

• Design has two distinct phases
• Phase 1, no outcome data available:
  – Propensity score analysis to approximate randomized block experiment for home hospital type
  – Ensure subclassification can create balance on covariates for large and small home hospital types
• Phase 2, uses intermediate outcome data on transfers:
  – Outline of the analysis for estimating causal effect of treating hospital type
  – Ensure within each subclass that there appear to be compliers who are treated in both and large and small treating hospitals
Figure 5.1: Cardia Cancer, Number of People, Subclassified by Propensity Score

Figure 5.2: Cardia Cancer, Difference in Means for Binary Covariates and Pscore

Figure 5.3: Cardia Cancer, t-statistics for Continuous Covariates

Outline of Analysis within Each Subclass

• Critical that we anticipate compliers in both large and small treating hospitals within each subclass

• Monotonicity assumption = no defiers; medically very plausible

\[
\text{ITT} = \pi_{LS} \text{ITT}_{LS} + \pi_{SS} \text{ITT}_{SS} + \pi_{LL} \text{ITT}_{LL}
\]

\[
CACE \equiv \text{ITT}_{LS} = \frac{1}{N_{LS}} \sum_{i \in LS} (Y_i(L) - Y_i(S)),
\]

\[
\text{ITT} = \pi_{LS} \text{ITT}_{LS}, \quad \text{ITT}_{LS} = \frac{\text{ITT}}{\pi_{LS}}.
\]
Method of Moments Estimates of the Number of Compliers Treated in Large and Small Hospital Types Under Monotonicity

<table>
<thead>
<tr>
<th>“Assigned”/Randomized Home Hospital Type</th>
<th>Treating Hospital Type</th>
<th>Approximate N in LS Principal Stratum Subclass</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>T</td>
<td>1</td>
</tr>
<tr>
<td>$\ell$</td>
<td>L</td>
<td>3</td>
</tr>
<tr>
<td>s</td>
<td>S</td>
<td>14</td>
</tr>
</tbody>
</table>
Table 5.3: Cardia Cancer: Observed Counts in Observed Groups and Approximate Counts in Principal Strata Under Monotonicity Assumption – Subclass 3

<table>
<thead>
<tr>
<th>“Assigned”/Randomized Home Hospital Type</th>
<th>Treating Hospital Type ( T )</th>
<th>( # )</th>
<th>Underlying Principal Strata: ( h= )</th>
<th>Approximate Proportion in Population in Principal Strata</th>
<th>Approximate N in ( LS ) Principal Stratum</th>
</tr>
</thead>
<tbody>
<tr>
<td>( h )</td>
<td>( \ell )</td>
<td>17</td>
<td>( \ell )</td>
<td>( s )</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( s )</td>
<td>( s )</td>
<td>13</td>
<td>( s )</td>
<td>( s )</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary: Objective Observational Study Design

• Should approximate a randomized experiment
  – No ultimate outcome data used or examined – “prospective”
  – Carefully consider decision-makers and the covariates used to make treatment assignments
  – If dataset is missing key covariates, usually do NOT continue
  – Use propensity score estimation to help create subclasses or matched pairs that achieve “balance” on covariates
  – Balance means treated and control subjects have distributions of covariates that are at least as similar as if they had been randomized into treatment and control
  – Analysis takes place within each subclass, and then answers are combined across subclasses