PSI conference
Design of a good comparative effectiveness study – “What does good look like?”

Mark Belger
Principal Research Scientist
Eli Lilly
The opinions expressed in this presentation and on the following slides are solely those of the presenter and not necessarily those of Eli Lilly.
Abstract

Following the launch of a new intervention, information is often required to understand how the new intervention is performing outside of the clinical trial setting. Comparative effectiveness studies using retrospective databases will answer some of these questions. However, these are often limited in the information they have available, and so prospective observational studies may be considered.

These post Launch prospective observational studies need careful planning and have a number of operational and methodological challenges. We focus on how to design such a study considering aspects such as sample size and representativeness of the cohort. We discuss the merits of different statistical approaches, such as propensity adjustment, for reducing the biases associated with observational studies. Finally, we will consider the types of statements that are acceptable when conducting comparative effectiveness analysis.
Study objectives

♦ As with any good study design a clear measurable primary objective is required. “Estimand”
  • Challenges
    – how to avoid study is perceived as being a seeding study
    – Interested in knowing about the newest intervention available
    – Is an observational study the best design for the question you want to answer
Rationale for Conducting an Observational Study

1. To address questions not easily obtained from clinical trials results (relevant for payers and physicians)
   i. Observational studies provide invaluable information on effectiveness, patient-reported outcomes and costs in a real-world environment – all important for the decision-making process

2. To address questions relating to the generalizability of clinical trial results to routine care (relevant for payers)

3. To obtain prospective data on new biologic treatment options coming to market for comparative effectiveness (relevant for Health Technology Assessment and payers)

4. To supplement disease registry data
   i. Disease registries are the cornerstone of safety data but focus much less on measures of effectiveness, particularly in some countries
   ii. Disease registries do not generally support access to comparator data to answer key scientific questions
Rationale for Conducting an Observational Study

There are Continuous Real World Data Needs During the Life-Cycle of a Drug

Development
- Post marketing commitments (safety etc.)
- Budget impact
- Unmet need / disease burden
- Patient recruitment
- Understand standard of care
- Trial design

Growth phase
- Adherence
- Utilization / prescribing patterns
- Long-term clinical outcomes

Mature phase
- Head to head comparative effectiveness
- Differentiation in sub-populations
- Target populations
- Usage Difference
- Effects of switching on outcomes
- Differentiate with or vs. protected galenics
- Competitor goes generic

Evidence required

Launch
- Conditional pricing review
- New competition
- New formulation / indication

Past
- New competition

Now

Target populations

Usage Difference

Effects of switching on outcomes

Differentiate with or vs. protected galenics

Competitor goes generic

Conditional pricing review

New competition

New formulation / indication

Long-term clinical outcomes

Head to head comparative effectiveness

Differentiation in sub-populations

Target populations

Usage Difference

Effects of switching on outcomes

Differentiate with or vs. protected galenics

Competitor goes generic

Conditional pricing review

New competition

New formulation / indication

Long-term clinical outcomes

Head to head comparative effectiveness

Differentiation in sub-populations

Target populations

Usage Difference

Effects of switching on outcomes

Differentiate with or vs. protected galenics

Competitor goes generic

Conditional pricing review

New competition

New formulation / indication

Long-term clinical outcomes

Head to head comparative effectiveness

Differentiation in sub-populations

Target populations

Usage Difference

Effects of switching on outcomes

Differentiate with or vs. protected galenics

Competitor goes generic

Conditional pricing review

New competition

New formulation / indication

Long-term clinical outcomes

Head to head comparative effectiveness

Differentiation in sub-populations

Target populations

Usage Difference

Effects of switching on outcomes

Differentiate with or vs. protected galenics

Competitor goes generic

Conditional pricing review

New competition

New formulation / indication

Long-term clinical outcomes

Head to head comparative effectiveness

Differentiation in sub-populations

Target populations

Usage Difference

Effects of switching on outcomes

Differentiate with or vs. protected galenics

Competitor goes generic
### Why a prospective observational study?

<table>
<thead>
<tr>
<th>Treatment patterns</th>
<th>Single-arm Treatment cohort</th>
<th>Multiple-arm/class cohort</th>
<th>Registry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-life use of biologic treatment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Use of concomitant medications</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Factors leading to treatment choice: ±

<table>
<thead>
<tr>
<th>Comparative Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinician rated outcomes</td>
</tr>
<tr>
<td>Patient reported outcomes (PROs)</td>
</tr>
<tr>
<td>HCRU/costs</td>
</tr>
<tr>
<td>Persistence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drug survival / Persistence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient profile</td>
</tr>
<tr>
<td>Persistence</td>
</tr>
<tr>
<td>Dosing regimen (mis)use</td>
</tr>
<tr>
<td>Patient Satisfaction</td>
</tr>
</tbody>
</table>
Sample size

- Based on precision not hypothesis testing
- What sample size provides optimal precision
- Budget considerations
- Enrolment
  - Market share of treatments
  - Drop out / endurance

![Sample Size Diagram](chart.png)
Statistical Challenges in Real World Data Comparative Effectiveness

- **With randomization** – standard methods produce estimates of causal treatment effects

- **Without randomization** – standard methods produce only ‘associations’ …. Treatment groups are NOT comparable at baseline thus comparisons are BIASED

**#1 Issue:** Confounding
The Observational Research Challenge

**Confounders**

- A variable is a Confounder if it is associated with both treatment selection and outcome
  
  - Example:
    - Risk Factors for Bleeding
    - Physicians avoid prescribing Trt A if risk factors present

**Measured**: Information is collected within the study and statistical adjustment is possible

**Unmeasured**: Information on the confounder is not available from the study
Adjustment for Baseline Confounding

- Restriction

- Regression (multi-variable)
  - Often same results as from propensity scoring (Shah 2005, Sturmer 2006)
  - Less robust than propensity scoring (Golinelli 2012, D’Agostino 2007, Baser 2007)
    - Large baseline difference between groups
    - Interactions/Nonlinear effects or rare events

- Propensity Scoring
  - (Rosenbaum & Rubin 1983)
  - Gold Standard

- Advanced Approaches
The Propensity Score (PS)

- PS – the conditional probability that a patient received treatment A given their set of observed baseline covariates X

- Usually computed via logistic regression

- Idea: compare treatments between patients with similar propensity scores allows “apples to apples” comparisons
  - Practical even when there are a large number of covariates to adjust for unlike direct stratification
Basic Assumptions for Causal Inference

Propensity Score adjustments can provide for estimates of the causal treatment differences under the following assumptions:

#1 No Unmeasured Confounders
All confounders are in the dataset and analysis

#2 Sufficient Overlap in Populations
positivity, no perfect confounding

#3 Correct Statistical Models
Basic Methods for Implementing PS

**Regression**
Simple regression model
\[ Y = \text{Trt} + \text{PS} \]

**Stratification**
Group patients with similar PS; Compare cohorts within each PS strata; then average across the strata

**Matching**
Match patients with similar PS, then compare Cohorts of matched pairs

**Inverse Weighting**
Run weighted analysis, weighting each patient by the inverse of their PS
Which PS Approach is Best??

Regression
Simple but least robust

Stratification
Often a good compromise between bias control and generalizability

Matching
Best for Bias Control – but can exclude significant numbers of patients

Inverse Weighting
Caution must carefully control highly influential values
Choosing the Propensity Model

- Variables to Include: Conflicting Advice
  - Be Inclusive  (Stuart 2010)
  - Be Restrictive  (Brookhardt 2006)
  - Use Data Driven Processes  (e.g. Schneeweiss 2009)
  - Stepwise Combination  (Imbens & Rubin 2015)
  - Include INTERACTIONS  (Imbens & Rubin 2015, Zagar 2017)

- Where All Agree
  - No variables that are affected by treatment  (Stuart 2010)
  - Finalize Model Without Outcome Data  (Rubin 2007, Xu 2017)
  - Check Overlap and Balance
    - Ensure overlap allows for analysis (Equipoise)
    - Ensure the model adequately balances the covariates
*Steps to a Quality Propensity Score Analyses

1. **Estimate the Propensity Score**
2. **Assess Overlap & Balance**
3. **Estimate the Treatment Effect**
4. **Sensitivity Analyses**

**Independent Analyst (Multistage: Xu 2017)**

**Design without outcome in sight (Rubin 2007)**
Propensity score distributions

For Summary Measures of Overlap see Imbens & Rubin (2015), Tipton (2014)

Stop: Extremely Strong Assumptions Require Causal Inference in Full Population
- “Preference Score” Guidance for when to even continue analysis
Assessing the Balance

• Hypothesis Testing
  • Common – but sample size dependent

• Standardized Differences
  • Recommended (Austin 2010)
  • “difference in means / pooled SD” (not sample size dependent)
  • Rule of Thumb: < 0.1 is OK
Improving Bias Control?

- **Matching**
  - Optimal, Exact, With Replacement

- **Balancing Algorithms**
  - Entropy: can produce exact balance in means and variances between groups

- **Tree Based**
  - Random forest; gradient boosting, local control, …

- **Regression**
  - Elastic Net, HDPS, …

What is best?

Simulations (Austin 2006; Zagar 2016): It depends! No best method across all scenarios
Frequentist Model Averaging (FMA)
(Zagar 2017)

FMA Estimate

Test Sample (2/3)
Validation Sample (1/3)

PS Strata
PS Matching
Prognostic
Elastic Net
Method n-1
Method n

Model Fitting & Cross-Validation

Weight 1
Weight 2
Weight 3
Weight 4
... Weight n-1
Weight n
*Unmeasured Confounding: Solutions*

- **None**
  - Plausibility
  - Adjusted Analysis
  - Falsification outcome
  - Pseudo Trt.
  - Partial identification
  - IV
  - Empirical Calibration
  - Difference in differences
  - Regression discontinuity

- **Internal**
  - Plausibility
  - Adjusted Analysis
  - R & R sensitivity
  - Rosenbaum Sensitivity
  - Multiple imputations
  - Propensity calibration

- **External**
  - Plausibility
  - Adjusted Analysis
  - R & R sensitivity
  - Rosenbaum Sensitivity
  - Bayesian Twin Regression
Recent Guidance Documents

2007 STROBE
- 22 Item Checklist

2009 ISPOR Good Res. Practices
- Design and Reporting (Berger et al); Mitigating Bias (Cox et al); Analytic Methods (Johnson et al)

2010 GRACE
- Dreyer et al (2010); ISPE

2014 PCORI & ISPOR-AMPC-NPC
- Methodology Reports; Flowchart (Berger et al 2014)

2017 Joint ISPOR-ISPE TaskForce
Any Questions?