Causal Modeling with Applications in Clinical & Policy Studies

Sharon-Lise Normand

Blavatnik Institute, Department of Health Care Policy, Harvard Medical School
Department of Biostatistics, Harvard Chan School of Public Health

Thanks: R01-GM11339

April 10, 2019
Introduction

- Regulatory Science
- What is Evidence?
- Examples

Common statistical issues
- Observational data - risk adjustment
- Clustering
- Multiple related outcomes* or multiple competing treatments

Concluding remarks
Regulatory Science

- Medical product entry
  - **Food & Drug Administration**: evaluates medical product safety, effectiveness, and quality
  - Post-market **safety** assessments

- Insurance coverage for medical products/services
  - **Medicare Evidence Development and Coverage Advisory Committee**: evaluates medical literature, technology assessments, etc., on benefits, harms, and appropriateness of medical items and services to make health care coverage recommendations
    - Must **extrapolate** treatment benefits to their population

- Transportation
  - **Federal Motor Carrier Safety Administration**: evaluates the safety of large trucks, buses, and commercial vehicles
    - Use data on safety-based regulations collected from **roadside** inspections and crash reports

Sometimes, must rely on **observational** data
Scientific Evidence for Medical Decisions

1. Introduction

The current paradigm for integrating clinical information in real-world settings to establish whether benefits outweigh risks is out-dated. Divergence from this paradigm involves the recognition that randomized controlled trials that often serve as the basis for new technology approval are small and short-term, and post-market studies are often voluntary and difficult to implement. These problems have become increasingly important over the last decade because technology is changing at a rapid pace, therapies are utilized outside their intended populations, and more representative groups of patients are likely to have differential responses to the same therapy.

2. How are decisions made?

When a new medical technology is approved or cleared for use by the U.S. Food and Drug Administration, the manufacturer of the technology typically follows with a request to the Center for Medicare and Medicaid Services (CMS) for either a local coverage decision or a national coverage decision. This is a decision on whether CMS will provide coverage for the new technology. The CMS is the U.S. federal agency that administers the Medicare, Medicaid, and States Childrens Health Insurance Programs, and as such, is the largest health insurer in the United States. The majority of requests for coverage by CMS are for local decisions because a denial resulting from a national request has much more serious consequences than does one from a local request. National coverage decisions made by the CMS have substantial impact on the use of new technologies because other insurers frequently follow suit.

In some cases, the CMS will seek guidance from the Medicare Evidence Development & Coverage Advisory Committee (MedCAC), an independent advisory committee which was established in 1998. The MedCAC judges the strength of the evidence that supports a recommendation for coverage.

3. Accumulation of information to support or refute a theory or hypothesis

Replication important

Underlying mechanism important

What is evidence?

Sharon-Lise T. Normand\textsuperscript{a,b,*} and Barbara J. McNeil\textsuperscript{a,c}

The assumption that comparative effectiveness research will provide timely, relevant evidence rests on changing the current framework for assembling evidence. In this commentary, we provide the background of how coverage decisions for new medical technologies are currently made in the United States. We focus on the statistical issues regarding how to use the ensemble of information for inferring comparative effectiveness. It is clear a paradigm shift in how clinical information is integrated in real-world settings to establish effectiveness is required. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: evidentiary evaluation; multiple outcomes and comparisons; hierarchical Bayesian models; Bayes factors; posterior predictive probability

1. Introduction

The current paradigm for integrating clinical information in real-world settings to establish whether benefits outweigh risks is out-dated. Divergence from this paradigm involves the recognition that randomized controlled trials that often serve as the basis for new technology approval are small and short-term, and post-market studies are often voluntary and difficult to implement. These problems have become increasingly important over the last decade because technology is changing at a rapid pace, therapies are utilized outside their intended populations, and more representative groups of patients are likely to have differential responses to the same therapy.

\begin{itemize}
  \item In this paper, Tunis and colleagues identify several areas of improvement to provide "more and better evidence of what works." Their thesis, and that of many others, relies on the premise that comparative effectiveness research will provide "timely, relevant evidence." While we agree with their general premise, the framework for determining what constitutes evidence and how it can be obtained in a timely manner motivates our commentary.
\end{itemize}

We begin with a review of how coverage decisions for new medical technologies are currently made in the United States. We then describe the statistical methods for aggregating the ensemble of information to make inferences.
Hierarchy of Evidence

Therefore, **many designs** contribute to evidence base
Contemporary Setting

With data acquisition technologies, biggest challenge is **data integration**

- molecular biology
- toxicology
- genotoxicology
- imaging
- functional MRI
- electronic health records
- mobile applications
DuMouchel and Harris. Bayes methods for combining the results of cancer studies in humans and other species. JASA 1983;78(382):293-308.

Combining heterogenous data sources not new
Schizophrenia

- High clinical need & difficulty navigating health care system
- Low quality & black/white disparities
- Medicaid is largest payer
  - Antipsychotics are costliest therapeutic category for state programs

Horvitz-Lennon et al., Health Services Research, 2014
Schizophrenia

- High clinical need & difficulty navigating health care system
- Low quality & black/white disparities
- Medicaid is largest payer
  - Antipsychotics are costliest therapeutic category for state programs

- Medicaid policies/benefits vary across states but within-state:
  - Covered by a single payment system with identical policies/benefit structure
  - Variation in access to federal, state, and local (county) resources
  - Extant evidence suggests geography plays a role

Horvitz-Lennon et al., Health Services Research, 2014
Introduction

Example 1

Schizophrenia

1. Does **quality of mental health care** differ among black, white, and Latino Medicaid beneficiaries?
2. Do quality and disparities **change** over **time**?
3. Do quality and disparities **vary** across **counties**?
Drug Reformulations

- Manufacturer reformulate existing products to extend product life cycle (1984 Hatch-Waxman Act)
- Shift demand for original formulation (soon lose patent protection) to the reformulation
- Reformulations involve less frequent dosing, gradual release of active ingredient, or easier to administer
- Antidepressant reformulations common (original vs reformulation):
  - Celexa vs Lexapro (single isomer); Paxil vs Paxil CR (controlled release); Remeron vs Remeron Soltab (disolvable tablet)
- Clinical trial evidence is sparse; mixed at best
- Huskamp et al., Health Affairs 2009
Drug Reformulations

Do anti-depressant reformulations decrease medication discontinuation rates compared to original formulations?
Specific Drug Eluting Stents

- Rapid proliferation of drug eluting stents (DES)
- U.S. has 2nd highest number of overall stent insertions per capita
- Multiple competing versions supported by a few manufacturers
- Differences include polymer coating, specific drug, platform type, and delivery system
- Study **21,000+** adults, **10** model-specific DES, **3** manufacturers
- Rose & Normand; Biometrics 2018
Specific Drug Eluting Stents

Do particular model-specific DES cause fewer adverse cardiac events compared to other model-specific DES?
Common Themes

- Observational data
- Clustered data
  - Adults living within counties
  - Patients nested within hospitals
- Multiple competing treatments: 10 different drug eluting stents
- Multiple outcomes: quality indicators
Lack of Randomization

- Causal inference
  - Special case of **predictive** inference among subjects who could receive any of the different treatment options
  - Renewed interest

- Why increased interest?
  - Increasing availability of data
  - Increasing availability of different data types (e.g., text, images, etc.)
  - Ignorable treatment assignment more plausible by conditioning on more data

- Notation
  - $T = 1$ new and $T = 0$ standard treatment (assume binary)
  - $Y$ observed outcome
  - $Y_1, Y_0$ **potential outcomes** under $T = 1$ and $T = 0$

- Assumption that potential outcome exists is fundamental
Causal Assumptions

1. Sample is representative of **target** population

2. Outcomes for one subject is independent of treatment assignment of other subjects **and** treatments are well-defined & the same for all subjects (SUTVA)

3. Within the subpopulations defined by the confounders, treatments are **randomly assigned**
   - Untestable assumption (sensitivity analysis, multiple comparison groups, control outcomes)

4. There are subjects from all treatment groups at every combination of observed confounders
   - **Structural** violations
   - Practical violations due to **finite** sample size
   - Statistically **testable**

5. Constant (vs Non-constant) treatment effect
## Assumptions: Randomized vs Not

<table>
<thead>
<tr>
<th>Feature</th>
<th>Clinical Trial</th>
<th>Observational Setting Study</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment, T*</td>
<td>Well-defined and does not vary across subjects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(T)</td>
<td>Known</td>
<td>Unknown</td>
<td>Estimate</td>
</tr>
<tr>
<td>Positivity</td>
<td>Yes</td>
<td>Not</td>
<td>Real &amp; finite sample</td>
</tr>
<tr>
<td>0 &lt; P(T) &lt; 1</td>
<td>by design</td>
<td>necessarily</td>
<td>sample violations</td>
</tr>
<tr>
<td>Comparability</td>
<td>Almost always</td>
<td>Sometimes</td>
<td>Assess balance on observables</td>
</tr>
<tr>
<td>Effect Estimate</td>
<td>Intention to Treat (ITT)</td>
<td>Adjusted</td>
<td>Sensitivity to measured and hidden</td>
</tr>
<tr>
<td></td>
<td></td>
<td>as-treated</td>
<td>confounders**</td>
</tr>
<tr>
<td>Postrandomization Bias</td>
<td>Possible</td>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Loss to follow-up, time-varying risks, competing risks</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Assume treatment received by subject A does not affect the outcome of subject B; **Mitigation strategies: falsification outcome, multiple control groups; Kunz, Rose, Spiegelman, Normand (Chapter 1); Hernán, Robins (Chapter 3); Methods in Comparative Effectiveness Research, 2017*
Joint Distribution

\[ P(Y, T, X) = P(Y | T, X) \times P(T | X) \times P(X) \]
\[ = Q_Y \times \Pi_T \times Q_X \]

1. Treatment effect depends **only** on \( Q_Y \) **and** \( Q_X \)

\[ E_X (E(Y | T = 1, X)) - E_X (E(Y | T = 0, X)) \]

2. \( \Pi_T \) is the **propensity score** (nuisance)

\[ \Pi_T = P(T = 1 | X) \]

\( X \) could be **very high-dimensional**
1. Model **only** the treatment assignment mechanism via regression
   - Propensity score, $\Pi_T = P(T \mid X)$
   - Weight or match using $\hat{\Pi}_T$

2. Model **only** the outcome via regression
   - Multiple regression modeling
   - Parametric g-computation
     - Step 1: Estimate regression model for outcome and treatment association
     - Step 2: Plug predictions from Step 1 into parameter mapping for causal parameter

3. Model **both** the treatment assignment mechanism and outcome
   - Augmented Inverse Probability of Treatment Weights
   - Target Maximum Likelihood
Reformulation Example

<table>
<thead>
<tr>
<th></th>
<th>July 1, 1999</th>
<th>June 30, 2001</th>
<th>April 1, 2002</th>
<th>March 31, 2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paxil</td>
<td></td>
<td>Paxil</td>
<td>Paxil CR</td>
<td>Paxil CR</td>
</tr>
<tr>
<td>Paxil CR</td>
<td></td>
<td></td>
<td>Paxil</td>
<td></td>
</tr>
</tbody>
</table>

Non-Contemporaneous

Contemporaneous
Reformulation Example

<table>
<thead>
<tr>
<th>July 1, 1999</th>
<th>June 30, 2001</th>
<th>April 1, 2002</th>
<th>March 31, 2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paxil</td>
<td>Paxil</td>
<td>Paxil CR</td>
<td>Paxil CR</td>
</tr>
</tbody>
</table>

Rx Dates
- 4/1/02-3/31/04 Paxil CR = 24224
- 7/1/99-6/30/01 Paxil = 36811

Non-Contemporaneous (Matched)

Rx Dates
- 4/1/02-3/31/04 Paxil CR = 23713
- 7/1/99-6/30/01 Paxil = 23713

Contemporaneous (Matched)

Rx Dates
- 4/1/02-3/31/04 Paxil CR = 14307
- 4/1/02-3/31/04 Paxil = 14307
## Reformulation Example

### Non-Contemporaneous Matched Samples†‡

**Reformulation** versus Original Formulation

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Days to Discontinue</th>
<th>RR† (95% CI)</th>
<th># Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexapro</strong></td>
<td>91</td>
<td>0.83 (0.80, 0.85)</td>
<td>18,045</td>
</tr>
<tr>
<td>Celexa</td>
<td>64</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td><strong>Paxil CR</strong></td>
<td>64</td>
<td>0.87 (0.85, 0.89)</td>
<td>23,713</td>
</tr>
<tr>
<td>Paxil</td>
<td>61</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td><strong>Remeron Soltab</strong></td>
<td>65</td>
<td>1.04 (1.00, 1.08)</td>
<td>10,820</td>
</tr>
<tr>
<td>Remeron</td>
<td>66</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

†Kaplan-Meier Analysis of Risk of Antidepressant Discontinuation
‡No change with **contemporaneous** matched sample except for Remeron Soltab; RR = 0.94 (0.90, 0.98)
Target Maximum Likelihood

- **Key idea**: no need to maximize entire likelihood \((Y, T, X)\) because causal parameter only depends on \(Q_Y\) and \(Q_X\)
- Step 1: Estimate outcome model and treatment model
- Step 2: Plug predictions from Step 1 into parameter mapping for causal parameter

Step 1:

\[
E^*(Y | T = t, X) = E^0(Y | T = t, X) + \epsilon_t H^*(T, X) \\
H^*(T, X) = \frac{T}{\Pi_T} - \frac{1 - T}{1 - \Pi_T}
\]

Step 2:

\[
\frac{1}{N} \sum_{i=1}^{N} (E^*(Y | T = 1, X_i) - E^*(Y | T = 0, X_i))
\]

- **\(E^*\)**: **targeted** estimate of regression of \(Y\) on \((T, X)\) obtained by moving the initial estimate \(E^0\) by fluctuations defined by \(\epsilon_t H^*(T, X)\)
## Summary

<table>
<thead>
<tr>
<th>Approach</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPTW</td>
<td>Simple</td>
<td>Large variance estimates</td>
</tr>
<tr>
<td>1/(\Pi_T)</td>
<td></td>
<td>Weight trimming bias</td>
</tr>
<tr>
<td>Regression</td>
<td>Parametric</td>
<td>Extrapolation</td>
</tr>
<tr>
<td></td>
<td>Simple</td>
<td>if violate positivity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Functional form</td>
</tr>
<tr>
<td>G-Comp</td>
<td>Parametric</td>
<td>Extrapolation</td>
</tr>
<tr>
<td></td>
<td>Simple</td>
<td>if violate positivity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Functional form</td>
</tr>
<tr>
<td>A-IPTW</td>
<td>Double robust</td>
<td>Finite sample inefficient</td>
</tr>
<tr>
<td></td>
<td>Asymptotic efficiency</td>
<td></td>
</tr>
<tr>
<td>TMLE</td>
<td>Double robust</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asymptotic efficiency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Finite sample efficiency</td>
<td></td>
</tr>
</tbody>
</table>
Clustering

- Clustered data: when units are nested completely within other units
  - Longitudinal data: observations are clustered within subjects
  - Levels of clustering may be $> 2$

- Problem introduced: observations within a common unit are statistically dependent

- In practice: the between-unit variance may be a nuisance parameter or it may be of interest

- Marginal models (GEE) treat the between-unit variation as a nuisance parameter
  - Regression parameters represent association of patient-level covariates with changes in the population mean outcome

- Hierarchical models (mixed models, random effect models) introduce random effects
  - Regression parameters represent association of patient-level covariates with changes in the patient’s outcome
Clustered data

Clustering: Other Considerations

- Interest in covariate effects at different levels of the hierarchy
- Cross-level interactions

\[
Y_{ij} \mid \beta_i = x_{ij} \beta_i + z_{ij} \alpha + \epsilon_{ij}; \epsilon_{ij} \sim N(0, \sigma_i^2)
\]
\[
\beta_i \mid \tau^2 = w_i \gamma + u_i; u_i \sim N(0, \tau^2)
\]

- \( w_i \) is a county-level covariate
- \( x_{ij} \) is subject \( j \) living in country \( i \) covariate

Can always derive marginal model from conditional model
Multiple Outcomes

- Increasingly collected in clinical trials to measure effectiveness or efficacy
  - Label extensions
- Often measures are **non-commensurate**
  - Measured on different scales
- Common approaches:
  - Consider each outcomes separately using a univariate framework
  - Create a composite measure
- A multivariate approach would:
  - Use information contained in the correlation between outcomes
  - Permit better control over Type I error rates
  - Answer intrinsically multivariate questions
Schizophrenia Example: Quality Indicators

**Established** quality indicators endorsed by professional societies

- Cover pharmacological, psychosocial, and appropriateness services
- Majority operationalizable using billing data:
  - adequate clozapine dose and duration
  - avoidance of antipsychotic polypharmacy
  - use of psychosocial services
  - follow-up within 7-days of hospital discharge
  - few emergency department visits
- At least **15** established measures of quality of care

Measures are non-commensurate but thought to reflect a single concept (quality of care)

- Why clustered within county?
  - Counties are the administrative unit
**Likelihood-based Approaches**

Key Idea: avoid **direct** specification of the multivariate likelihood.

- **Factorization (a few variables)**

- **Introduction of a latent variable to model correlation among the multiple outcomes**

- **Quasi-likelihood: use quadratic exponential model to develop joint estimating equations**
Remarks

- If strong correlation between outcomes, univariate approach results in less efficient estimates.
- Higher efficiency gains realized when the two outcomes share different sets of covariates.
- When missing data, can directly maximize the likelihood for the latent variable approach owing to conditional independence assumption.
- Many situations with more than two outcomes:
  - Latent variable approach is easily extended to several continuous and binary outcomes; not true for factorization approach.
Concluding Remarks

- **Risk adjustment for causal inference**
  - Must assess validity of assumptions
  - Must assess robustness of conclusions to reasonable departures from assumptions
  - High-dimensional data: parametric assumptions difficult to verify and more uncertainty

- **Clustering**
  - Understand question being posed
  - Is variability a nuisance parameter or is it of policy-interest?

- **Multiple outcomes**
  - Are outcomes manifest variables to inform about a latent variable
  - Do covariates affect outcomes differently
Schizophrenia Example: County Disparities

California

Worse Care vs Whites

Better Cares vs Whites

Florida

New York

North Carolina

(Minority - White) Quality Score

(Minority - White) Quality Score

(Minority - White) Quality Score

(Minority - White) Quality Score
Key Principles

- Avoid **strong** parametric assumptions
  - In setting with many confounders, very likely to get the model really wrong
- Adhere to causal inference **assumptions**
  - Validate assumptions
  - Assess robustness to reasonable departures from assumptions
- Adopt a **design-based** approach
  - Separate treatment from outcome during the modeling process
- Reflect all **uncertainty** in estimates

Thank you