What is the Value of Big Data in Comparative Effectiveness Research and Clinical Decision Making?

May 18, 2015

Issues Panel

Moderator:  William H. Crown, PhD  
Chief Scientific Officer, Optum Labs

Panelists:  Sarah Greene, MPH  
Associate Director, CER Methods and Infrastructure Program, Patient-Centered Outcomes Research Institute (PCORI)

Miguel Hernán, MD, DrPH  
Professor of Epidemiology and Biostatistics, Harvard T. H. Chan School of Public Health

Milton C. Weinstein, PhD  
Henry J Kaiser Professor of Health Policy and Management, Harvard T. H. Chan School of Public Health
Market Context

Velocity

Tests and Treatments
(Medical, Lab, Pharmacy Claims, Standardized Costs)
Health Risk Assessments
Socioeconomic
(Race, Income, Education, Language, ...)
Vital Signs
Medication Orders
Admissions, Discharges, Transfers
Patient Health Survey (PHQ-9)
Health Survey Measurement (SF-12, SF-36)

Care Coaching Engagements
Evidence Based Medicine
(Recommended Care Pathways)
Mobile Applications / Social Networking
Medical Research
Genomic

Future

Volume

Tests and Treatments
(Medical, Lab, Pharmacy Claims, Standardized Costs)
Health Risk Assessments
Socioeconomic
(Race, Income, Education, Language, ...)
Vital Signs
Medication Orders
Admissions, Discharges, Transfers
Patient Health Survey (PHQ-9)
Health Survey Measurement (SF-12, SF-36)

Care Coaching Engagements
Evidence Based Medicine
(Recommended Care Pathways)
Mobile Applications / Social Networking
Medical Research
Genomic

Strengths of Claims, Clinical and Linked Data

Claims Data

- **Benefits**
  - Comprehensiveness of medical services and Rx at the patient level (cross-provider)
  - Retail and specialty Rx across settings
  - Cost data
  - 20 years of expertise in its use and design

- **Limitations**
  - Inaccuracies in diagnosis coding
  - Lack of clinical outcomes and clinical severity measures

Clinical Data

- **Benefits**
  - All payer, including self-pay
  - Clinical data/severity measures/problem lists/lab results
  - Rich data extractable from clinical notes via NLP
  - Patient-specific data shows all services received within the specific practice/IDN

- **Limitations**
  - Unknown range of patient care experience that is represented in EMR data (without claims data)

Linked Data

- **Benefits**
  - Feasible to link claims and EMR without exchanging PHI using salting and hashing methods
  - Richness of clinical supplement to the comprehensiveness of claims
  - Able to address and measure inherent bias
  - Clinical endpoints for outcomes studies (especially safety)
  - Support analysis related to primary non-adherence
  - Imputation techniques may help fill missing data gaps
Natural Language Processing (NLP) often needed to extract information and convert into structured data

<table>
<thead>
<tr>
<th>Types of notes analyzed</th>
<th>Examples of extracted data elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Outpatient office visits</td>
<td><strong>History</strong></td>
</tr>
<tr>
<td>• Consultant notes</td>
<td>• Nausea, cough, fatigue, diarrhea, dizziness</td>
</tr>
<tr>
<td>• Operative (procedure) reports</td>
<td>• Pain level</td>
</tr>
<tr>
<td>• Admission notes (often with H&amp;P)</td>
<td>• Home HbA1c measurements</td>
</tr>
<tr>
<td>• Discharge summaries</td>
<td>• Cancer stage</td>
</tr>
<tr>
<td>• Nursing notes</td>
<td>• Smoking status</td>
</tr>
<tr>
<td>• Emergency room notes</td>
<td>• Diabetes family history</td>
</tr>
<tr>
<td>• Pathology notes</td>
<td>• Exercise amounts</td>
</tr>
<tr>
<td>• Radiology notes</td>
<td><strong>Physical</strong></td>
</tr>
<tr>
<td>• Cardiology notes</td>
<td>• Physical exam findings</td>
</tr>
<tr>
<td></td>
<td>• Height, weight, BMI</td>
</tr>
<tr>
<td></td>
<td>• Vitals: BP, pulse, temperature</td>
</tr>
<tr>
<td></td>
<td>• Review of symptoms</td>
</tr>
<tr>
<td></td>
<td><strong>Labs</strong></td>
</tr>
<tr>
<td></td>
<td>• Ejection fraction</td>
</tr>
<tr>
<td></td>
<td>• PFTs (FEV1, FVC)</td>
</tr>
<tr>
<td></td>
<td>• Path: HER2, tumor margins, levels</td>
</tr>
<tr>
<td></td>
<td>• Bone density</td>
</tr>
<tr>
<td></td>
<td><strong>Medications including OTC</strong></td>
</tr>
<tr>
<td></td>
<td>• Including strength, route, frequency, form</td>
</tr>
<tr>
<td></td>
<td><strong>Other</strong></td>
</tr>
<tr>
<td></td>
<td>• EDSS</td>
</tr>
<tr>
<td></td>
<td>• Glasgow Coma Scale</td>
</tr>
<tr>
<td></td>
<td>• Physician reason for change of medication</td>
</tr>
</tbody>
</table>

Power is in data linkage

• Linking EHR and claims data addresses many of the limitations of each source alone.

• Challenges in protecting patient privacy and complying with privacy legislation.
  – Need ability to generate de-identified linkage variables
  – Notes are inherently identifiable so need NLP to extract content.
  – Even more severe challenges will arise with respect to genetic and genomic information

• None of this considers the patient voice. PROs, like EHRs, are a critical data element for many conditions but are less widely available. PRO collection needs to be integrated into health plan operations and clinical care pathways in provider settings before such data will be widely available.

• In the meantime, we will need to continue to rely upon modeling methods to bring different data types together. Imputation methods may also be possible.
Implications

• Can data linkage get us to the Holy Grail of causal inferences with observational data? If so, under what conditions?

• Data science methods such as machine learning techniques tend to focus on prediction rather than estimation of treatment effects. What is the role for these methods?

• What about the voice of the consumer? The patient perspective is generally missing in claims and EMR databases.

• Some situations may be too complex to model directly—nonlinearities, feedback loops, etc. in healthcare systems. What methods work best in these situations? In what ways does big data help?

• QALYs in health economic models are tied to the concept of consumer utility but utility data is not generally available in traditional claims databases. What does that imply for estimating consumer demand with such data?

• Optimization methods from operations research have not been widely applied to examine outcomes research questions. Is this an opportunity?

The Intersection of Large Data Capabilities, Outcomes Research, and Decision-Making

Sarah Greene, MPH
Associate Director, CER Methods & Infrastructure, PCORI
“The Why” for Medicine: Our national clinical research system is well-intentioned but flawed

• High percentage of decisions not supported by evidence*
• Health outcomes and disparities are not improving
• Current system is great except:
  – Too slow, too expensive, and not reliable
  – Doesn’t answer questions that matter most to patients
  – Unattractive to clinicians & administrators
  – *Has yet to harness immense potential of health data*


“The Why” for Research: What’s the big deal about a reusable infrastructure for CER?

• Often, clinical trials coordinating centers and large research networks are established for one purpose or one funding cycle
• New trial or new condition typically begets construction of a new infrastructure
• PCORnet blends the capabilities of healthcare systems and patient-driven organizations in support of a sustainable national ecosystem for research that is more efficient and patient-centered than our current system
• Tackling inefficiencies in the current research process, with a heavy injection of patient engagement and massive data resources
“The Why” for Data

“One great strength of prospective research remains the fact that data needs can be identified in advance and collected according to rigorous, pre-specified, and validated standards. Routinely collected patient data rarely meets such standards. **Most patient data has been collected to serve immediate clinical and business needs, not for research purposes.** Often there is significant variation in the categorization of data, the structure of reported data, and also the methods of soliciting and recording data.

---

**Big Data Attributes: Relevance to Research**

- **Big Data is Typified by the “3 V’s”**
  - *Volume:* Unparalleled amount of data available...
  - *Velocity:* Accumulating at incredibly fast rate...
  - *Variety:* And it differs from source to source

- **Some other “V” words emerging**
  - *Veracity:* Are the data correct and true, can we draw reliable conclusions?

  *Value!* Do the data produce something that someone wants and needs?
In the Big Data Era Consider “P” as well as “V”

- **Privacy**: Establishing conditions under which data are shared for research
- **Protection**: Ensuring that the data holders and data users apply appropriate technical, operational and physical safeguards
- **Patient-Centeredness**: Involving patients in decisions about how the data may be used
  - Data use preferences of a rare disease patient advocacy group may be very different from those sitting on the patient advisory council of an academic medical center

Our intent for PCORnet:
3Vs + 3Ps will equal 3Rs

- Capitalize on the **volume**, **velocity** and **variety** of data sources
  
  AND...

- Take a **patient-centered** approach to data **privacy** and **protection**
  
  THEN...

- With those building blocks, PCORnet’s use of big data will be: **Reliable, Rigorous, and Research-Ready**
**PCORnet Common Data Model – Why It Matters for CER**

- Variety is Typical of EHR Data!...
  - Variety of Encounter Types, EHR Vendors, Data Entry Modules, Coding Systems, Users, etc.
- Getting data out for use in research—especially multi-site research—takes effort, time, expertise, and money
- Pace of research can be drastically improved if we can efficiently move data out of EHRs and map them into a common model
- Mapping local data to a common format means that if all sites store their local data in that format, one research data extraction program can be written and run against that data, averting re-work at each site
## Glycosylated hemoglobin (HBA1c) units

<table>
<thead>
<tr>
<th>%</th>
<th>%T.HGB</th>
<th>% TL HGB</th>
<th>% HGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEMOGLOBIN</td>
<td>%T.Hgb</td>
<td>% OF TOTAL</td>
<td>PERCENT</td>
</tr>
<tr>
<td>U</td>
<td>%T.Hgb</td>
<td>% of Hgb</td>
<td>Percent</td>
</tr>
<tr>
<td>%HB</td>
<td>% NGSP</td>
<td>% of total</td>
<td>HbA1c%</td>
</tr>
<tr>
<td>% OF T</td>
<td>%NGSP</td>
<td>%THb</td>
<td>%HbA1c</td>
</tr>
<tr>
<td>%AIC</td>
<td>% TOTAL HGB</td>
<td>%NGSP</td>
<td>% A1C</td>
</tr>
<tr>
<td>MG/DL</td>
<td>G/DL</td>
<td>mmol/mol</td>
<td>Blank</td>
</tr>
<tr>
<td>% AIC</td>
<td>% Alc</td>
<td>%Hb</td>
<td>g/dL</td>
</tr>
<tr>
<td>NULL</td>
<td>%THb</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All 34 values are valid, but create work when trying to derive data to compare across patients, settings, providers, episodes

Raebel Pharmacoepi and Drug Safety 2014; DOI:10.1002

### Data Visualization: After 7th refresh, partner A
Data Visualization: After 8th refresh, partner A

New data problem in old time period

Data Visualization: After 8th refresh fixed
Utility of Big Data is in Thoughtful Application

- What do decision-makers need?
- What do they value?
- Can we apply big/new data sources with appropriate rigor?
- How can we produce answers at the pace and scale needed for healthcare decision-making, with its shorter time horizons?
- How can we place our research results in the larger context of health care improvement?
PCORnet as a Big Data Resource

- If it succeeds, PCORnet will be able to leverage:
  - Health system information (diagnoses, procedures, EHR)
  - Mobile health data (FitBit, Ginger.io, symptom trackers)
  - Biologic data from stored blood, saliva, tissue
  - Online patient communities & social media

- Not turn-key or automatic!
  - Determining veracity and validity of data, AND transforming the data into research-ready form is requires resources and expertise

- Privacy and security controls are paramount
  - PCORnet is partnering with patients to devise data sharing policies and practices

- Art of finding ways for data to simultaneously support healthcare research and healthcare improvement

Using Big Data to emulate a target trial when a randomized trial is not available

Miguel Hernán
DEPARTMENTS OF EPIDEMIOLOGY AND BIOSTATISTICS

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH
We need to make decisions NOW

- Treat with A or with B? Treat now or later? When to switch to C?

- A relevant randomized trial would, in principle, answer each comparative effectiveness and safety question
  - Interference/scaling up issues aside

But we rarely have randomized trials

- expensive, untimely, unethical, impractical

- And deferring decisions is not an option
  - no decision is a decision: “Keep status quo”

- **Question:**
  - What do we do if we have Big Data?
Answer:
We conduct observational studies

- but only because we cannot conduct a randomized trial

- Observational studies are **not** our preferred choice
  - For each observational study, we can imagine a hypothetical randomized trial that we would prefer to conduct
  - If only it were possible

The **target trial**

- An observational study in a **large health care database** can be viewed as an attempt to emulate a hypothetical, nonblinded randomized trial

- If the observational study succeeds at emulating the target trial, both studies would yield identical effect estimates
  - except for random variability
Procedure to answer each clinical/policy question:

☐ Step #1
  ■ Describe the protocol of the target trial

☐ Step #2
  ■ Option A: Conduct the target trial
  ■ Option B
    □ Use observational (Big) data to explicitly emulate the target trial
    □ Apply appropriate causal inference techniques and Big Data analytics

Key elements of the protocol of the target trial

☐ Eligibility criteria
☐ Treatment strategies
  ■ randomly assigned at start of follow-up
☐ Start/End of follow-up
☐ Outcomes
☐ Causal contrast(s) of interest
☐ Analysis plan
The observational study needs to emulate

- Eligibility criteria
- Treatment strategies
  - randomly assigned at start of follow-up
- Start/End of follow-up
- Outcomes
- Causal contrast(s) of interest
- Analysis plan

EXAMPLE
Epoetin dosing and mortality

- Question: What is the effect of different doses of epoetin therapy on the mortality risk of patients undergoing hemodialysis?

- Data: US Renal Data System (Medicare claims database)
  - ~18,000 eligible elderly patients
  - Zhang et al. CJASN 2009; 21:638-644
The target trial

- Eligibility criteria
  - End-stage renal disease

- Strategies
  - Fixed weekly dose of intravenous epoetin
  - 15,000, 30,000, or 45,000 units

- Follow-up
  - From 3 months after hemodialysis onset until death, loss to follow-up or administrative end of the study (1 year)

- Outcome
  - All-cause mortality

Methodological challenge

- Time-varying treatment
  - Use and dose of epoetin varies over the course of the disease

- Time-varying confounders
  - Hematocrit level, comorbidities
  - may be affected by prior treatment

- Treatment-confounder feedback
  - Need “causal” methods
  - Inverse probability weighting of marginal structural models
Survival under 3 epoetin dosing regimes

Zhang et al. CJASN 2009; 21:638-644

But this is a silly target trial

- In clinical practice, patients do not receive a fixed weekly dose of epoetin
  - That would be clinical malpractice
- Rather, actual clinical strategies are dynamic
  - A patient’s weekly dose depends on her hemoglobin or hematocrit, which in turn depends on her prior weekly dose
More reasonable strategies for a target trial

1. Mid-Hematocrit strategy
   - epoetin to maintain Hct between 34.5% and 39.0%

2. Low-Hematocrit strategy
   - epoetin to maintain Hct between 30.0% and 34.5%.

- Under both strategies, epoetin dose is
  - increased by >10% if previous Hct below target
  - decreased by <10% times [previous Hct minus lower end of range] or increased by <10% times [upper end of range minus Hct] if Hct within target
  - decreased by >25% if Hct above target

More reasonable strategies imply more work

- Need to specify a more detailed protocol for the target trial

- Need to specify how to emulate that protocol
  - Appropriate adjustment for time-varying confounders becomes critical
  - Zhang et al. Medical Care 2014
Survival under these 2 dynamic strategies

<table>
<thead>
<tr>
<th>Components</th>
<th>Hypothetical</th>
<th>Cumulative, Unlimited</th>
<th>Randomized Clinical Trial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>To study the risks and benefits of adjunct therapy to target treatment (34.5% vs 30.0%)</td>
<td>Same</td>
<td>Same</td>
</tr>
<tr>
<td>Study population</td>
<td>Initially 1,500 patients with diabetes aged 65 or older, randomized to either intervention or usual care; participants monitored for 6 months.</td>
<td>Same</td>
<td>Same</td>
</tr>
<tr>
<td>Eligibility criteria</td>
<td>Age &gt; 65 years, HbA1c &lt; 7%, no history of cardiovascular disease</td>
<td>Same</td>
<td>Same</td>
</tr>
<tr>
<td>Follow-up</td>
<td>Start after completing 30 days of lifestyle therapy</td>
<td>Same</td>
<td>Same</td>
</tr>
</tbody>
</table>

**Survival under these 2 dynamic strategies**

- Low Hot
- Mid Hot

5/18/2015

Hernan - Target trial
Advantage of the target trial approach for Big Data analysis

- Facilitates the comparison of complex strategies that are sustained over time and may depend on a patient’s evolving characteristics
  - Dynamic treatment strategies
  
  - Not “treat vs no treat” but rather “when to treat, when to switch, when to monitor” depending on time-varying factors

What Is The Role of Cost-Effectiveness Models in a World of “Big Data”?

Milton C. Weinstein

Department of Health Policy and Management
Comparative Effectiveness Research Initiative
Harvard T.H. Chan School of Public Health
Problems with Observational Health Data

- People who get different interventions (treatments, tests) may not be comparable to each other

  - Some predictors of treatment choice are observable

  - But some are not observable

  - Some are affected by treatment and vary over time

- You can’t measure what didn’t happen

  - Counterfactual interventions or clinical strategies that weren’t used

  - New interventions
Decision-Analytic Approach

• Specify a clinical starting point
• Identify two or more clinical strategies
• Assign probabilities to health outcomes
• Value health outcomes according to preferences of affected persons
• Evaluate costs of each strategy
• Compare strategies
• Do sensitivity analyses

Clinical starting point

• Important clinical characteristics may not be identified
  – Symptoms
  – Test results
  – Response to previous treatments
Clinical strategies

• You can’t always tell from data what clinical strategy a patient is getting
  – Dependence of intervention on treatment response, side effects, prior test results

• Some clinical strategies may not be represented
  – Innovative starting or switching criteria
  – New treatments and diagnostics

Probabilities of Health Outcomes

• Long-term data may be unavailable
  – Lifetime data are almost never available

• Patients getting different strategies are different
  – Need causal inference models for valid inferences
Value Health Outcomes

• Real-time HRQoL data may not be available

• Patient health states for utility assignment may not be identifiable

• Utility data may not be available to link to discernable patient-level health states

Evaluate Costs

• Historically a strength of claims data

• Challenges:
  
  – Patients getting different strategies are different
  
  – Some cost drivers may not be observable
    • e.g., “soft” clinical trial endpoints such as symptoms
What Is The Role of Cost-Effectiveness Models in a World of “Big Data”?

• “Big Data” cannot provide all the inputs required for determining optimal clinical decisions or assessing cost-effectiveness

But...

• “Big Data” can be helpful in populating decision models

Discussion / Questions