Highlights from research in the Observational Medical Outcomes Partnership

Paul Stang, PhD
Patrick Ryan, PhD
Global Epidemiology
Janssen Research and Development

Public-Private Research Partnership established to inform the appropriate use of observational healthcare databases for studying the effects of medical products:
- Conducting methodological research to empirically evaluate the performance of various analytical methods on their ability to identify true associations and avoid false findings
- Developing tools and capabilities for transforming, characterizing, and analyzing disparate data sources across the health care delivery spectrum
- Establishing a shared resource so that the broader research community can collaboratively advance the science

Why Most Databases Were Created
- Payors track expenditures to manage costs
  - Managed care
  - Insurance company
  - Employer
  - Government - some track more than health care
- Clinicians track patients to manage care
  - Mayo Clinic, GPRD, Electronic Medical Records
- Surveys are carried out in the population
  - National Comorbidity Survey, National Household Survey (UK), Whitehall, MONICA, Framingham
- The data reflect the underlying healthcare delivery system
- Most were not designed for research
Types of Health Care Databases

- Claims data: MD, goods, prescriptions
- Electronic medical record
- Laboratory/diagnostic testing
- Symptom data: eg, NHANES
- Facility: eg, Hospital data (Premier)
- Cost database: Prescription AWP
- Surveillance data: SEER, CDC, AERS, MAUDE
- Prospective cohorts: Framingham
- Prescription data: Medco, IMS
- Cross-sectional studies: NC3
- Registries: Cancer, pregnancy, disease, exposure, devices
- Linked National Databases: Sweden, Denmark

To understand the data, you must understand....

- Underlying healthcare delivery system
- How ‘long’ and ‘deep’ the capture is for a given patient
- Limitations of coding that may affect interpretation
  - Financial incentives in coding in some countries
  - ‘Rule-out’ diagnoses
- How different methods perform in different databases and populations

What can we do with longitudinal observational databases?

- Disease natural history
  - What is the background rate of a disease?
  - What are risk factors for developing the disease?
  - Does having the disease impact risk of subsequent outcomes?
- Drug utilization patterns
  - How many patients have exposure to medical product?
  - Amongst patients with exposure, what are common comorbidities and concomitant medications?
- Exploring the effects of medical products
  - Safety signal detection and evaluation
  - Comparative effectiveness research
Prepare data once for research, use it multiple times for many purposes

- Medical product safety surveillance (drugs, vaccines, devices)
- Comparative effectiveness research
- Health economics
- Quality of care

[Link to CDM vocab]

Patient profile: longitudinal record with diabetes

A brief summary from a long journey

All Symposium materials (presentations, handouts, references, and even YouTube videos) are all publicly available at:

[Link to Symposium Presentations]
OMOP 2010/2011 Research Experiment

- 10 data sources
- Claims and EHRs
- 200M+ lives
- Open-source
- Standards-based
- 14 methods
- Epidemiology designs
- Statistical approaches adapted for longitudinal data
- OMOP Methods Library
- Inception cohort
- Case control
- Logistic regression
- Common Data Model

Drug-outcome pairs

Criteria for positive controls:
- Event listed in Boxed Warning or Warnings/Precautions section of active FDA structured product label
- Drug listed as "causative agent" in Tisdale et al, 2010: "Drug-Induced Diseases"
- Literature review identified no powered studies with refuting evidence of effect

Criteria for negative controls:
- Event not listed anywhere in any section of active FDA structured product label
- Drug not listed as 'causative agent' in Tisdale et al, 2010: "Drug-Induced Diseases"
- Literature review identified no powered studies with evidence of potential positive association

OMOP 2011/2012 Research Agenda

Drug-outcome pairs

Methods development

- Evaluate study design (EDDIE)
- Methods enhancements
  - Multivariate self-controlled case series
  - Increased parameterization
  - Case-control, new user cohort designs
  - Application of existing tools
  - ICTPD, OS, LGPS, DP
- Expand CDM for additional use cases

OMOP Ground truth for OMOP 2011/2012 experiments

<table>
<thead>
<tr>
<th></th>
<th>Positive controls</th>
<th>Negative controls</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoniazid</td>
<td>27</td>
<td>17</td>
<td>44</td>
</tr>
<tr>
<td>Fluticasone</td>
<td>23</td>
<td>17</td>
<td>40</td>
</tr>
<tr>
<td>Indomethacin</td>
<td>23</td>
<td>17</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
<td>51</td>
<td>124</td>
</tr>
</tbody>
</table>

Criteria for positive controls:
- Event listed in Boxed Warning or Warnings/Precautions section of active FDA structured product label
- Drug listed as "causative agent" in Tisdale et al, 2010: "Drug-Induced Diseases"
- Literature review identified no powered studies with refuting evidence of effect

Criteria for negative controls:
- Event not listed anywhere in any section of active FDA structured product label
- Drug not listed as 'causative agent' in Tisdale et al, 2010: "Drug-Induced Diseases"
- Literature review identified no powered studies with evidence of potential positive association
Exploring isoniazid and acute liver injury

Smith et al. 2011 study design and results

- Data source: Administrative claims from health insurance board of Quebec
- Study design: Cohort
- Exposure: all patients dispensed >=30d of therapy, 180d washout
- Unexposed cohort: 2 patients per exposed, matched by age, gender, and region, with no tuberculosis therapy
- Time-at-risk: Length of exposure + 60 days
- Events: Incident hospital admission for noninfectious or toxic hepatitis
- "Event ratio" estimated with conditional logistic regression
- Covariates: prior hospitalization, Charlson score, comorbidities

Revisiting the isoniazid – acute liver injury example

- Data source: MarketScan Medicare Beneficiaries (MDCR)
- Study design: Cohort
- Exposure: all patients dispensed new use of isoniazid, 180d washout
- Unexposed cohort: Patient with indicated diagnosis (e.g. pulmonary tuberculosis) but no exposure to isoniazid; negative control drug referents
- Time-at-risk: Length of exposure + 30 days, censored at incident events
- Covariates: age, sex, index year, Charlson score, number of prior visits, all prior medications, all comorbidities, all priority procedures
- "Odds ratio" estimated through propensity score stratification (20 strata)
Observational Medical Outcomes Partnership

Receiver Operating Characteristic (ROC) curve

- ROC plots sensitivity vs. false positive rate
- Rank-order all pairs by RR from largest to smallest
- Calculate sensitivity and specificity at all possible RR thresholds
- Area under ROC curve (AUC) provides probability that method will score a randomly chosen true positive drug-outcome pair higher than a random unrelated drug-outcome pair
- AUC=1 is perfect predictive model
- AUC=0.50 is random guessing (diagonal line)
- Cohort method on MDCR: AUC = 0.64

Setting thresholds from an ROC curve

- If target sensitivity = 50%: RR Threshold = 1.25
  Sensitivity = 69% Specificity = 90%
- If threshold set to RR=2:
  Sensitivity = 4%
  Specificity = 98%
- Cohort method on MDCR: AUC = 0.64
  AUC suggests that this method is modestly predictive, on the low end of diagnostic tests used in clinical practice, but at any given threshold there is a high false positive rate and/or false negative rate
- Question: what strategies can be applied to do even better?

Strategies to improve predictive accuracy

- Stratify results by outcome
- Tailor analysis to outcome
- Restrict to sufficient sample size
- Optimize analysis to the data source
Performance after applying these strategies

To recap the improvements that could be achieved by following these ideas...

Before: One method applied to all test cases

<table>
<thead>
<tr>
<th>Outcome</th>
<th>AUC</th>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.64</td>
<td>1.25</td>
<td>69%</td>
<td></td>
</tr>
</tbody>
</table>

After: Partitioning, tailoring, restriction

<table>
<thead>
<tr>
<th>Outcome</th>
<th>AUC</th>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute kidney injury</td>
<td>0.92</td>
<td>2.69</td>
<td>95%</td>
<td></td>
</tr>
<tr>
<td>Acute liver injury</td>
<td>0.76</td>
<td>1.51</td>
<td>89%</td>
<td></td>
</tr>
<tr>
<td>Acute myocardial infarction</td>
<td>0.84</td>
<td>1.59</td>
<td>92%</td>
<td></td>
</tr>
<tr>
<td>GI bleed</td>
<td>0.86</td>
<td>1.87</td>
<td>94%</td>
<td></td>
</tr>
</tbody>
</table>

Optimal methods (AUC) by outcome and data source

<table>
<thead>
<tr>
<th>Data source</th>
<th>Acute kidney injury</th>
<th>Acute liver injury</th>
<th>Acute myocardial infarction</th>
<th>GI bleed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDCR</td>
<td>OS: 401002 (0.92)</td>
<td>OS: 401002 (0.76)</td>
<td>OS: 407002 (0.84)</td>
<td>OS: 402002 (0.86)</td>
</tr>
<tr>
<td>CCAE</td>
<td>OS: 404002 (0.89)</td>
<td>OS: 403002 (0.79)</td>
<td>OS: 408013 (0.85)</td>
<td>SCCS: 1931010 (0.82)</td>
</tr>
<tr>
<td>MDCD</td>
<td>OS: 408013 (0.82)</td>
<td>OS: 409013 (0.77)</td>
<td>OS: 407004 (0.80)</td>
<td>OS: 401004 (0.87)</td>
</tr>
<tr>
<td>MSLR</td>
<td>SCCS: 1939009 (1.00)</td>
<td>OS: 406002 (0.84)</td>
<td>OS: 403002 (0.80)</td>
<td>OS: 403002 (0.83)</td>
</tr>
<tr>
<td>GE</td>
<td>SCCS: 1949010 (0.94)</td>
<td>OS: 409002 (0.87)</td>
<td>KTP: 301601 (0.84)</td>
<td>SCCS: 3034001 (0.89)</td>
</tr>
</tbody>
</table>

- Self-controlled designs are optimal across all outcomes and all sources, but the specific settings are different in each scenario
- Acute kidney injury has consistently lower predictive accuracy
**Takeaways from research toward risk identification**

- **Performance of different methods**
  - Self-controlling
  - Evaluating
    - Broader definitions have better coverage and comparable performance to more specific definitions
- **Performance across different signal sizes**
  - A risk identification effect with
  - Data source heterogeneity
    - Substantial value
  - Method parameter sensitivity
    - Each method than other

**Consider a typical observational database study:**

*Exploring clopidogrel and upper gastrointestinal bleeding*

- **Error** = distance from the point estimate to the true effect
  - How far away from truth is RR=2.07?
- **Bias** = expected value of the error distribution
  - When applying this type of analysis to this type of data for this type of outcome, how far on average is the estimate from the true value?
- **Coverage** = probability that true effect is contained within confidence interval
  - When applying this type of analysis to this type of data for this type of outcome, do the 95% confidence intervals (1.66 to 2.58 in this case) actually contain the true relative risk 95% of the time?
Learning from what's already known

Prespecified Falsification End Points
Can TheyValidate True Observational Associations?

• Their recommendation: Use 3-4 negative controls, in addition to target outcome, as a means of assessing the plausibility of an observational analysis result
• Our recommendation: Use a large sample of negative (and positive) controls to empirically measure analysis operating characteristics and use them to calibrate your study finding

OMOP approach to methodological research

• Develop a standardized implementation of the analysis strategy
  – Study design: Case-control
  – Nesting within indication (unstable angina)
  – Case definition: First episode of upper GI hemorrhage
  – Exposure definition: Length of exposure + 30d
  – Exclusion criteria: <180d of observation before case
• Systematically apply the analysis across a network of databases, consistently for a large sample of positive and negative controls
  – GI Bleeding: 24 positive controls, 67 negative controls
  – Criteria for negative controls:
    • Event not listed anywhere in any section of active FDA structured product label
    • Drug not listed as 'causative agent' in Tisdale et al, 2010: "Drug-Induced Diseases"
    • Literature review identified no evidence of potential positive association
• Record all effect estimates (RR, CI) from all analysis-database-drug-outcome combinations and summarize analysis*database performance
  – If we assume drugs identified as negative controls truly have no effect on outcome, then we can assume true RR = 1 as a basis for measuring error

Case-control estimates for GI bleed negative controls

• If 95% confidence interval was properly calibrated, then 95%*65 = 62 of the estimates should cover RR = 1.
• We observed 29 of negative controls did cover RR = 1.
• Estimated coverage probability = 29 / 65 = 45%
• Positive tendency: 74% of estimates have RR > 1.
• Error distribution demonstrates positive bias (expected value > 1) and substantial variability.
Intuition for empirical calibration: You can use empirical null to adjust original estimate by 'shifting' for bias and 'stretching' for variance in error distribution at each true effect size.

**p-value calibration plot**

CC: 2000314, CCAE, GI Bleed

<table>
<thead>
<tr>
<th>p-value</th>
<th>calibrated p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; .05</td>
<td>55%</td>
</tr>
<tr>
<td>&lt; .05</td>
<td>6%</td>
</tr>
</tbody>
</table>

Clopidogrel - Bleeding:

Uncalibrated:

RR CI (1.79-1.93)

p < .001

Calibrated:

RR CI (0.79-4.57)

p = .30

**Interpreting observational studies: why empirical calibration is needed to correct p-values**

Figure 4. Effect estimates extracted from MEDLINE abstracts of observational studies using healthcare databases, by publication year. The number of estimates reaching statistical significance (p < .05) is estimated using four assumptions on the error distribution: null (traditional significance testing), mean = 0, SD = 1; small bias, mean = 0, SD = 0.25; moderate bias, mean = 0.25, SD = 0.25; large bias, mean = 0.5, SD = 0.5.
Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System

Takeaways from research toward risk identification

• Systematic exploration of negative and positive controls can be used to measure analysis operating characteristics of a signal detection system
  – Errors in observational studies were observed to be differential by design, data source, and outcome
  – Magnitude and direction of bias varied, but all analyses had error distributions far from nominal
  – OMOP results suggest a risk identification system can perform at AUC>0.80
    • Define potential for useful information, but...
    • Imperfect predictions mean substantial risk of false positive and false negative findings
• Traditional interpretation of p-values and 95% confidence intervals may be misleading in the context of observational database studies
• Empirical calibration is one approach to attempt to account for residual error that should be expected within any observational analysis
• Advancing the science of observational research requires a transparent, reproducible approach to methodology and systematic application

Lessons for building a risk identification system

• Strategies to improve performance:
  – Partition results by outcome
  – Tailor analysis to outcome
  – Restrict to sufficient sample size
  – Optimize analysis to the data source
• OMOP’s experimental evidence suggests that following these strategies may yield predictive accuracy at or better than most clinical screening tools used in standard practice
Lessons for building a risk identification system

- Where we are now:
  - Given the diversity in performance and heterogeneity in estimates, we caution against generalizing these results to other outcomes or other data sources.
  - If you want to apply risk identification to different outcomes and/or different data sources, we suggest performing an empirical assessment to establish best practice and benchmark performance.

- Potential next step:
  - Conduct similar experiment for additional 19 outcomes identified by EUADR\(^1\) as high-priority safety issues.
  - Once 23 HOIs complete, re-assess whether patterns emerge that would allow generalization to other outcomes.

\(^1\)Trifiro et al, PDS 2009

Conclusions

- Using the OMOP approach, a risk identification system can perform at AUC>0.80.
- Traditional p-values and confidence intervals require empirical calibration to account for bias in observational studies.
- Advancing the science of observational research requires an empirical and reproducible approach to methodology and systematic application.

Contact

Paul Stang
pstang@its.jnj.com

Patrick Ryan
pryan4@its.jnj.com

OMOP website: http://omop.org

Look for supplement in Drug Safety, published in Nov2013

OMOP 2013 Symposium, November 5-6, 2013