Optimal methods approaches in Existing Data: Lessons from OMOP and OHDSI communities

Patrick Ryan, PhD
Janssen Research and Development
Columbia University Medical Center
19 October 2016
April 2016: “The risk of the most serious cardiovascular events (cardiac ischemia or stroke) was not elevated in AI-only users compared with tamoxifen users”

Sept 2016: “Aromatase inhibitors are associated with a higher risk of MI compared with tamoxifen”
Non-steroidal anti-inflammatory drugs and myocardial infarctions: comparative systematic review of evidence from observational studies and randomised controlled trials

P A Scott, G H Kingsley, C M Smith, E H Choy, D L Scott


Figure 1 Analysis of 14 case-control and six cohort studies of myocardial infarction in users and non-users of non-steroidal anti-inflammatory drugs. n, events; N, subjects.
How do you *judge* what is a ‘good’ observational analysis?

- Validity of pre-specified protocol?
- Choice of study design?
- Selection of database?
- Quality of research team?
- Precision of confidence interval?
How could we **objectively measure** the reliability of observational evidence?

- **Consistency**: to what extent are the results robust to study design choices?
  - Reproducibility and external validity in multiple databases
  - Stability of exposure/outcome/covariate phenotype definitions
  - Statistical analysis parameter sensitivity

- **Discrimination**: to what extent can the analysis distinguish between null effects and positive effects?
  - Area under ROC curve (AUC) for estimates of negative and positive controls
  - Sensitivity/specificity-positive predictive value at decision thresholds

- **Calibration**: to what extent are the estimated statistics consistent with ground truth?
  - Systematic error distribution to evaluate magnitude of bias for effect estimates
  - Empirical null distribution using negative controls to determine if p-value truly represents probability of estimate when true RR = 1
  - Using positive controls to measure coverage probability to assess if 95% confidence interval actually contains the true effect size 95% of the time
A Comparison of the Empirical Performance of Methods for a Risk Identification System

Patrick B. Ryan · Paul E. Stang · J. Marc Overhage · Marc A. Suchard · Abraham G. Hartzema · William DuMouchel · Christian G. Reich · Martijn J. Schuemie · David Madigan

- Systematically applied multiple methods (cohort, case-control, self-controlled designs) to multiple databases (claims, EHR) for a large set of positive and negative control drug-outcome pairs
- Measured performance of each method in each database for each health outcome of interest (including AMI)
  - Consistency: database heterogeneity, parameter sensitivity
  - Discrimination: AUC
  - Calibration: systematic error, coverage probability
Performance of analysis varies by database and outcome, but self-controlled designs showed highest discrimination for AMI.
Lessons from the OMOP experiments

1. Database heterogeneity:
   Holding analysis constant, different data may yield different estimates

   “Evaluating the Impact of Database Heterogeneity on Observational Study Results”

2. Parameter sensitivity:
   Holding data constant, different analytic design choices may yield different estimates

   Madigan D, Ryan PB, Scheumie MJ, Therapeutic Advances in Drug Safety, 2013: “Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies”

3. Empirical performance:
   Most observational methods do not have nominal statistical operating characteristics

   Ryan PB, Stang PE, Overhage JM et al, Drug Safety, 2013:
   “A Comparison of the Empirical Performance of Methods for a Risk Identification System”

4. Empirical calibration can help restore interpretation of study findings

   Schuemie MJ, Ryan PB, DuMouchel W, et al, Statistics in Medicine, 2013:
   “Interpreting observational studies: why empirical calibration is needed to correct p-values”
OHDSI best practices for population-level effect estimation

Evidence Generation
- Write and share protocol
- Open source study code
- Use validated software
- Replicate across databases

Evidence Evaluation
- Produce standard diagnostics
- Include negative controls
- Create positive controls
- Calibrate confidence interval and p-value

Evidence Dissemination
- Don’t provide only the effect estimate
- Also share protocol, study code, diagnostics and evaluation
- Produce evidence at scale
OHDSI’s collaborative journey to reliable evidence generation

**Methodological research**
- Data quality assessment
- Common Data Model evaluation
- ATHENA for standardized vocabularies

**Open-source analytics development**
- WhiteRabbit for CDM ETL
- Usagi for code mapping
- Vocabulary exploration
- Database profiling

**Clinical applications**
- CohortMethod
- SelfControlledCaseSeries
- SelfControlledCohort
- TemporalPatternDiscovery
- PatientLevelPrediction
- APHRODITE for predictive phenotyping
- LAERTES for evidence synthesis
- PENELOPE for patient-centered product labeling
- Chronic disease therapy pathways
- Large-scale causality assessment

---

**Analytical use case**

**What is an ‘optimal’ method for population-level effect estimation in observational data?**

Any approach that is transparent, fully reproducible, and empirically demonstrated across multiple databases to discriminate between true effects with unbiased and well-calibrated estimates.
Join the journey

http://ohdsi.org