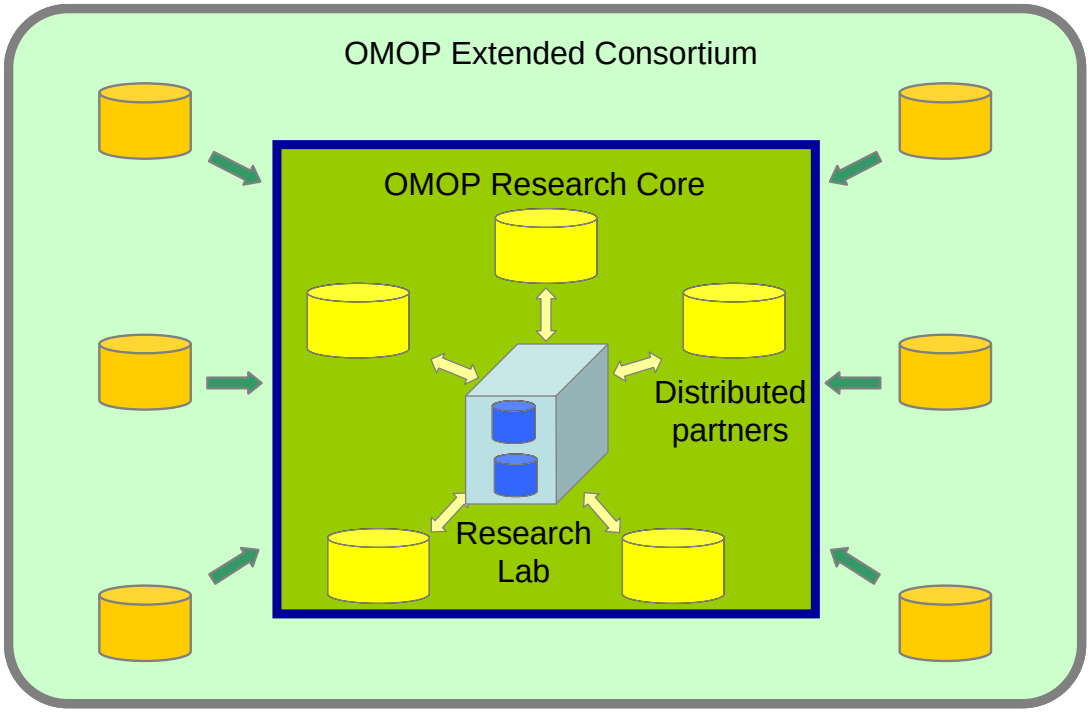
# Combining results across databases in a distributed network: an overview

Jesse A. Berlin, ScD
VP, Epidemiology
Johnson & Johnson Pharmaceutical R&D

**Statistical and Epidemiological Issues in Active Medical Product Surveillance**

The Brookings Institution • Washington, DC
February 16, 2011

# Overview of Partnership Design (OMOP)




- **OMOP Research Core** is responsible for designing, developing and managing the execution of the approved research proposals.
- **OMOP Research Lab** will be used to manage analysis process across all data sources within the Research Core.
- **Distributed Partners** implement the OMOP Common Data Model and execute protocols within their data environment
- The broader scientific community can participate in the **OMOP Extended Consortium**

# What are we estimating?

- Drug-event association(s)
  - Summary estimate of odds ratio, relative risk, or hazard ratio
- Assume this will be focused on a specific, pre-defined question, e.g.,
  - **REPORT TO FDA ON A PROTOCOL FOR ACTIVE SURVEILLANCE OF ACUTE MYOCARDIAL INFARCTION IN ASSOCIATION WITH USE OF A PHARMACEUTICAL AGENT (saxagliptin)**
  - (but I'm not sure it matters for this discussion)

# One option in the distributed network setting

- Generate separate estimates from all sites (partners)
- Present separate estimates, discuss the strengths and limitations of each estimate
- Don't combine estimates across sites
  - Argue that it's "not meaningful" (especially for small effects)
  - Shapiro S. Meta-analysis/Shmeta-analysis. Am J Epidemiol. 140(9):771-8, 1994 Nov 1.
- Choose an estimate (or not) based on:
  - Some (predefined) measure of internal validity?
  - Personal bias and preconceived idea (your own Bayesian prior)

# Some limitations of administrative databases

- Missing data on confounders:  smoking, OTC drugs, alcohol consumption
- Loss to follow-up (turnover in health care plans – an issue in the US)
- Data quality issues (e.g., miscoding)
- Limited effectiveness data (capture mostly clinical events, not symptoms)
  - Emphasis is on risk in the benefit-risk discussion

# Not combining could be a problem

- If we know, *a priori*, that some individual databases are poorly suited to the question at hand (or in general), then why use those databases to begin with?
- Risk of "cherry picking"
  - We can always find a rationale for excluding a result we don't like (or even results we do like) or selecting an estimate we do like
- The perils of the narrative review are what led to the widespread use of meta-analysis of clinical trials

# Options for combining across sites

- Combine adjusted estimates using inverse-variance weighted averages (of some kind)
  - Do the same adjustment within each database
  - Relies on common data model and common definitions and coding of confounders
  - Method has limitations when events are rare
  - Must decide on fixed vs. random effects, Bayesian, or other models
  - Worry about heterogeneity of estimates across sites

# More options

- Use individual-level data and do a “single estimate” (sharing all data with the central site)
  - This is essentially the same as the meta-analysis, except for the choice of weights
  - Privacy and proprietary issues probably preclude doing this, however
- Model maybe (?) more stable than IV-weighted average, especially with rare events

# A middle ground?

- Jeremy Rassen will discuss later

# Individual patient vs. aggregate-level data

- Generally, advantages include the ability to:
  - Adjust for patient-level covariates in a common way across all datasets
  - Do proper time-to-event analyses
  - Do proper subgroup analyses for patient-level characteristics
- Example references (forgive the self-citation):


- Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman, HI. Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Statistic in Medicine, 2002; 21:371-387
- Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level factors. J Clinical Epidemiology, 2004; 57:683-697

# Heterogeneity

- Variability among study-specific (site-specific) estimates of relative risk
- Can be incorporated into variance estimates for the combined estimate (RE) or not (FE)
- Potential sources of heterogeneity SHOULD ALWAYS be explored
  - Definitions of exposure and outcome (we can fix this part)
  - Other aspects of data (e.g., availability of lab data)
  - True clinical / biological variability (subgroups that may be differentially represented in different databases)
  - Other characteristics of the populations (databases)

# Conclusions

- Access to individual patient data (IPD) is crucial
  - Control of confounding
  - Subgroup analyses
  - Proper time-to-event analyses
- Privacy (and proprietary) issues generally preclude sharing detailed IPD centrally
  - So creating a single database and doing a single (still stratified) analysis is likely to be infeasible

# Conclusions (2)

- Conducting analyses at the sites is a practical alternative and should give the same answer
  - Depends on having a common data model that accommodates various "platforms" (not just SAS datasets)
- Rare events impose additional constraints at the site level and probably require a different approach
  - Data may be too sparse to generate stable results at the site level
- Worry about understanding sources of heterogeneity across sites
  - May relate to idiosyncrasies of site-specific data
  - May vary with the question

# Conclusions (3)

- We need to learn more, e.g.,
  - Ongoing OMOP work
  - PCORI methods committee
  - Others?
- “Be careful out there”
  - (Desk sergeant in the US television series, “Hill Street Blues”)