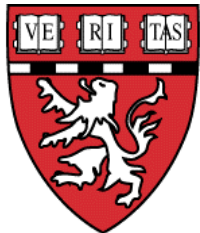


Multi-Institutional Studies Using Observational Data: Opportunities and Challenges

Jeffrey Brown, PhD

The 32nd annual Midwest Biopharmaceutical Statistics Workshop
Ball State University, Muncie, IN
May 18 – 20, 2009



Department of Ambulatory Care and Prevention
Harvard Medical School and Harvard Pilgrim Health Care





Outline

- Why multi-institutional studies
- Issues to consider
- Example: Meningococcal Vaccine Study
- Opportunities and Challenges

Why Multi-Institutional Studies?

- Large observational healthcare databases are valuable for studies of drug and vaccine safety, other public health questions
- Very large populations needed to address questions where
 - Exposure is rare
 - Outcome (or condition) is rare
 - Need for answer in limited timeframe
- Drug and vaccine safety issues often meet one or more of those criteria

Outline

- Why multi-institutional studies
- **Issues to consider**
- Example: Meningococcal Vaccine Study
- Challenges and opportunities



Multi-Institutional Studies

Key Questions

- What data are needed?
- Where will the data be stored?
- How will the data be analyzed?

What Data Are Needed?

- Population-based?
- Sample size?
- Claims (administrative, enrollment, billing, dx, px, rx)
- EMR (clinical information)
- Claims plus EMR
- Linkage to registry information (birth, tumor, death)
- Laboratory results
- Free-text notes
- Exposure & outcome verification via chart review?
- Access to treatments
- Access to information (complete capture of care)
- Longitudinal or cross-sectional
- Local collaborators

Where Are Data Stored?

Centralized vs. Distributed

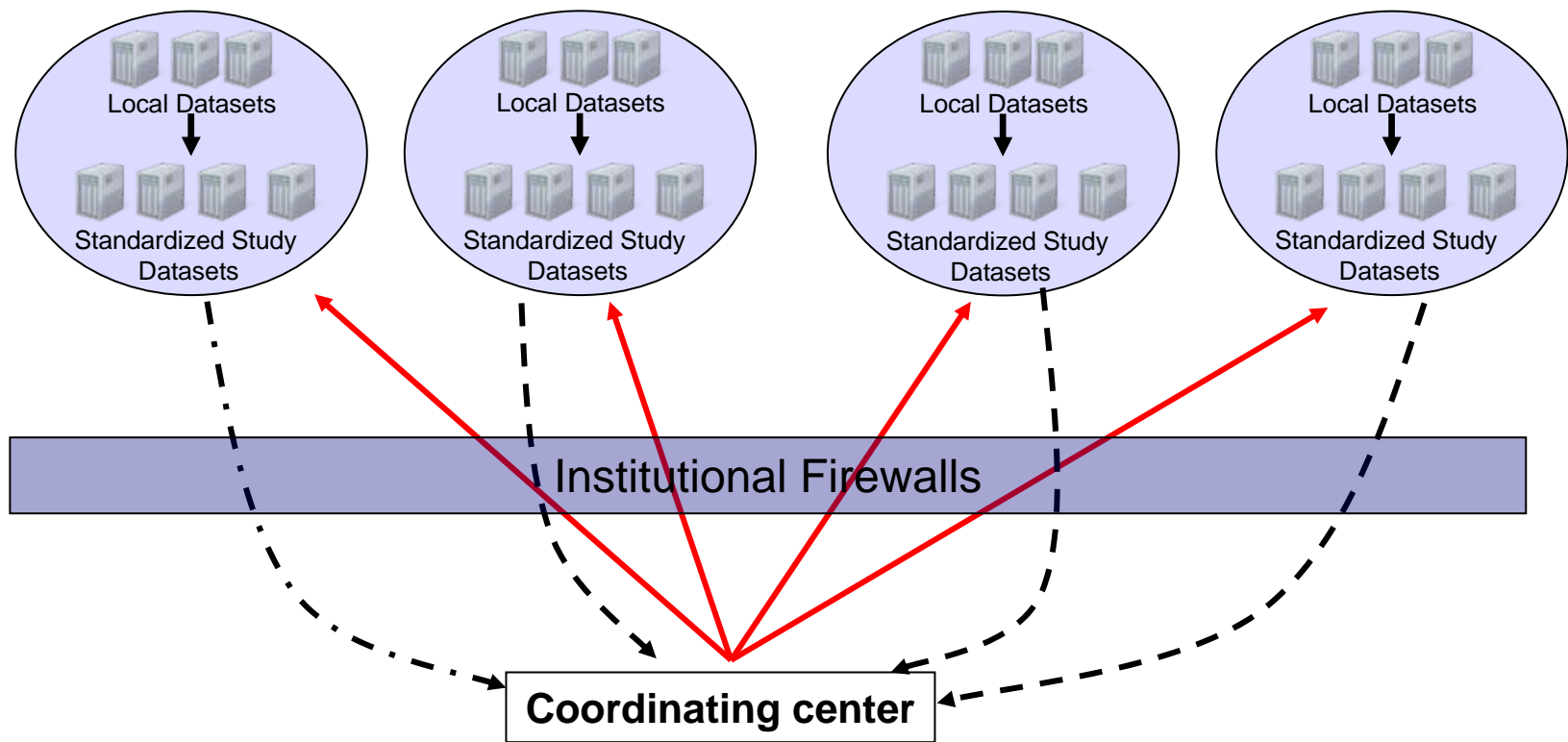
- Centralized: data owners send their data to a central location
- Distributed: data owners maintain physical control of their data, protected by their security rules
- Mixed model: studies that require analytic datasets from disparate sources to be analyzed together (e.g., multivariate analyses)

My Vote: Distributed

How Are Data Analyzed?

- Centralized data model: query or analysis run by anyone with data warehouse access and permission to execute the query
- Distributed data model: two options for querying/ analysis:
 - Independent analytics: Create a study protocol, ask each data owner to independently implement it locally, return results for aggregation or meta-analysis
 - No common data model required
 - Coordinated analytics: Create a study protocol, study team creates analytic code, distributes it to each data owner to run against the data they have stored in the *common model*

My Vote: Coordinated Analytics



Principles of distributed data model:

- Data remain under local control, no central data warehouse.
- Transfer as little data as possible; if stratified summary data, e.g., person-time and events, suffices, only that is transferred for combined analyses.
- If pooled person-level analyses are required, only transfer required fields and maximally de-identify, e.g., use offset dates, age ranges.

Analytic steps:

- Plans extract and transform local data to standard datasets
- Data quality checking led by coordinating center, using standard programs
- Analysis team sends analytic program to each site to run against study datasets.
- Sites run the program and return summary results or limited datasets.

Outline

- Why multi-institutional studies
- Issues to consider
- **Examples**
 - **Meningococcal Vaccine Study**
- Challenges and opportunities



The Meningococcal Vaccine Study Collaboration*

*Velentgas P, Bohn RL, Brown JS, Chan KA, Gladowski P, Holick CN, et al. A distributed research network model for post-marketing safety studies: the Meningococcal Vaccine Study. *Pharmacoepidemiol Drug Saf.* 2008;17(12):1226-34.

Meningococcal Vaccine Study

Background

- New meningitis vaccine (MCV4, Menactra) licensed January 2005
- Several case reports of GBS following Menactra vaccination
- **Goal:** Design and conduct a study with adequate power to answer question of increased risk of GBS following MCV4 vaccination
- **Challenges**
 - GBS is a rare adverse event
 - Prevalence of MCV4 vaccination is low overall (<10% in 11-18 year old population)
 - Need to validate outcomes with chart review
 - Controversial topic → requires substantial consideration of conflicts, priorities, reporting, governance
- **Conclusion**
 - Study requires combined efforts of several large health plans with active health research divisions
 - Requires input from a broad array of stakeholders

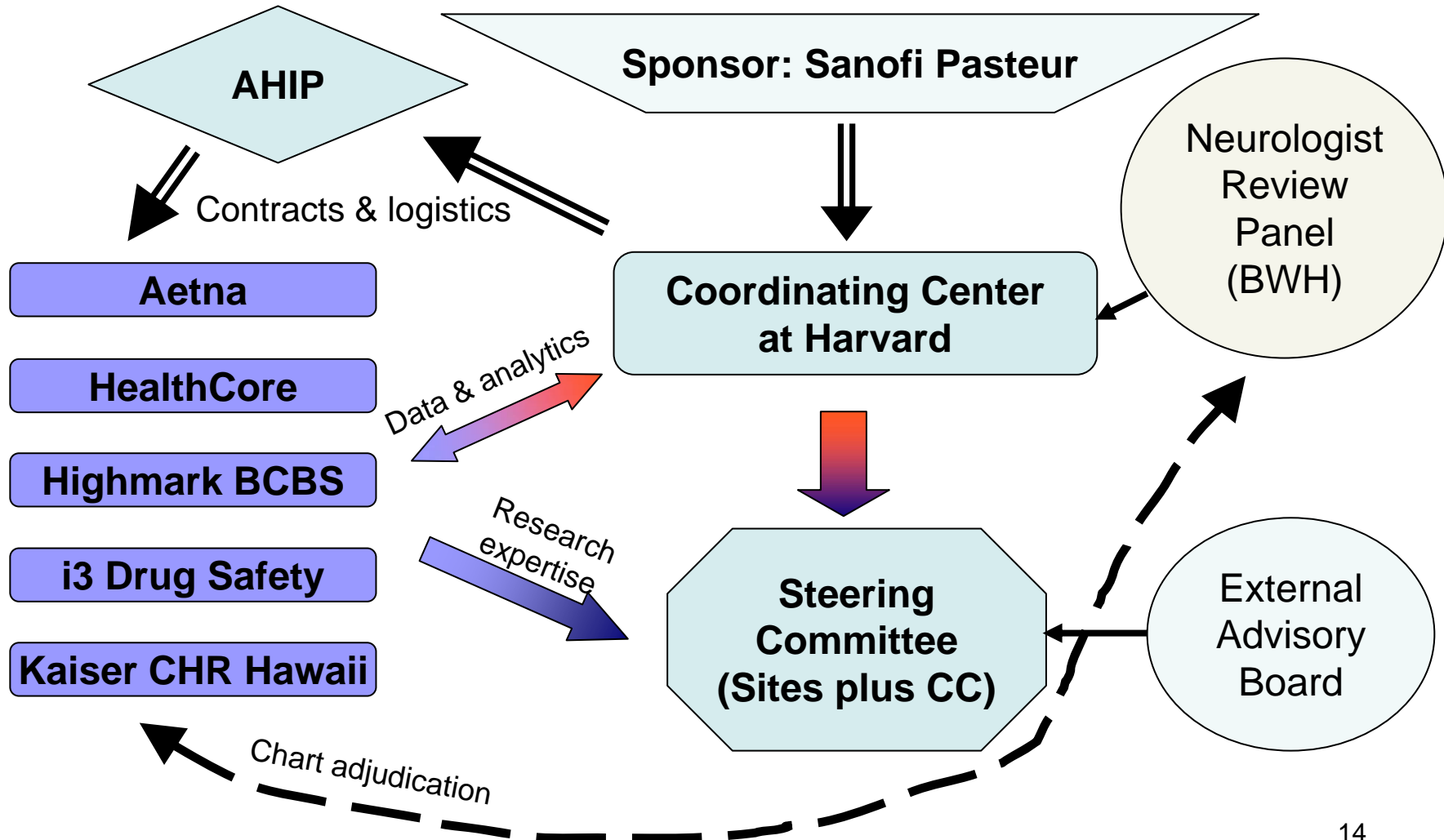
Meningococcal Vaccine Study

Design

- Multi-site retrospective cohort study (new collaborations)
- Collaboration between Harvard Coordinating Center and research arms of five US health plans
 - Aetna
 - HealthCore, Inc.
 - Highmark BCBS
 - Ingenix/ i3 Drug Safety
 - Kaiser Permanente Center for Health Research Hawaii
- Multiple interim data extractions and reports
- “Hybrid” cohort with nested case-control design
- Claims-based GBS cases confirmed through medical chart review and adjudication by neurologist panel

Registered at [ClinicalTrials.gov](https://clinicaltrials.gov) (NCT00575653)

Meningococcal Vaccine Study Structure



Some MVS Challenges

- Establishing Collaborations
 - Engaging with leaders of research organizations affiliated with large health plans → expertise, not just data, is required
 - Contracting, IRB
 - Trust, communication
- Governance
 - Scientific decision making
 - Study structure (Coordinating Center, EAC, etc)
- Analytic Approach
 - Develop acceptable common model
 - Data checking
 - Issue resolution
 - Trust with sensitive information



Multi-Institutional Observational Studies Challenges

(examples of what can go wrong)

Creating Study Datasets

“Enhancements” and new old data

- Protocol “enhancements”
 - Variation in procedure rates: *“We exclude claims if there is no clinician encounter on the same day”*
 - Too much (little) data included
- Process “enhancements”
 - Bonus exclusions: *“There was no provider reported on the claim so we excluded it from the list for chart review”*
- Extract consistency
- New old data

Creating Study Datasets

Data Checking Example

Data QC1: Demographics

	Site A	Site B	Site C	Site D	Total
Number of Records	31,518	54,252	170,365	124,705	380,840
Number of Patients	31,518	54,252	170,365	124,705	380,840
Sex					
Female	20,403 65%	34,094 63%	106,282 62%	77,974 63%	238,753 63%
Male	11,111 35%	20,158 37%	64,068 38%	46,717 37%	142,054 37%
Other/Missing	4 0%	0 0%	15 0%	14 0%	33 0%
Total	31,518	54,252	170,365	124,705	380,840
Age 2009					
0-4	34 0%	55 0%	317 0%	255 0%	661 0%
5-9	101 0%	321 1%	716 0%	734 1%	1,872 0%
10-19	1,132 4%	2,714 5%	6,035 4%	5,661 5%	15,542 4%
20-39	5,873 19%	12,165 22%	30,335 18%	21,526 17%	69,899 18%
40-64	15,952 51%	29,860 55%	81,804 48%	58,139 47%	185,755 49%
65-74	3,695 12%	5,219 10%	23,454 14%	18,840 15%	51,208 13%
75-99	4,688 15%	3,891 7%	27,547 16%	19,476 16%	55,602 15%
100-119	41 0%	27 0%	157 0%	74 0%	299 0%
Other/ Missing	2 0%	0 0%	0 0%	0 0%	2 0%
Total	31,518	54,252	170,365	124,705	380,840
Range of DOB					
Min DOB	1892.05	1905.01	1896.04	1898.12	
Max DOB	2008.10	2008.03	2008.11	2008.07	
mean	53.7	48.9	54.3	54.0	48.9 - 54.3

Creating Study Datasets

Data Checking Example

Data QC1: Procedure

	Site A	Site B	Site C	Site D	Total
Number of Records	5,354,897	13,108,382	18,762,527	6,198,190	43,423,996
Number of Patients	30,895	53,991	403,946	120,019	608,851
Avg Px per Patient	173	243	46	52	71
Encounter Type: Records					
AV	3,844,945 72%	6,786,218 52%	16,829,425 90%	5,100,796 82%	32,561,386 75%
ED	186,139 3%	935,130 7%	#N/A	602,739 10%	1,724,008 4%
IP	428,795 8%	717,638 5%	1,933,102 10%	494,655 8%	3,574,190 8%
IS	46,501 1%	258,436 2%	#N/A	#N/A	304,937 1%
LO	497,996 9%	1,040,657 8%	#N/A	#N/A	1,538,653 4%
OE	164,980 3%	3,160,090 24%	#N/A	#N/A	3,325,070 8%
RO	185,541 3%	210,104 2%	#N/A	#N/A	395,645 1%
TE	#N/A	109 0%	#N/A	#N/A	109 0%
Total	5,354,897	13,108,382	18,762,527	6,198,190	43,423,998
Encounter Type: Patients w/					
AV	30,787 100%	53,677 99%	167,661 42%	118,090 98%	370,217 61%
ED	17,770 58%	36,743 68%	#N/A	58,227 49%	112,740 19%
IP	15,257 49%	24,005 44%	343,315 85%	61,725 51%	444,304 73%
IS	3,041 10%	4,913 9%	#N/A	#N/A	7,954 1%
LO	19,386 63%	43,115 80%	#N/A	#N/A	62,501 10%
OE	11,139 36%	50,664 94%	#N/A	#N/A	61,803 10%
RO	22,149 72%	33,094 61%	#N/A	#N/A	55,243 9%
TE	#N/A	62 0%	#N/A	#N/A	62 0%
Date Range					
Min	2002.01	2000.03	1985.08	2002.01	range 1985.08 - 2002.01
Max	2008.12	2008.12	2009.03	2008.12	2008.12 - 2009.03

Creating Study Datasets

Data Checking

- We only need data for 2002 forward. **I will delete data prior to 2002**
- Do you have ED visits for dx, procedure, and events? Would it be possible to add? ED visit are part of the outcome algorithm. **We do, I will separate out ED visits from outpatient visit**
- The number of dx and procedures in the inpt setting seem high? Do you have any theories why? **We allow up to 13 dx and 8 px.**
- It looks like you only pulled pharmacy dispensing for AEDs. We will need **all** pharmacy dispensings for the cohort. **I will pull all rx data for AED users**
- There may be some extraneous patients in the procedure file that are not in the demographic file. Might be worth investigating. **I will delete those people.**

Creating Study Datasets

Data Checking

- Some demographic patients are not in the enrollment file. Non-enrolled patients may present for services and when they do they are included in the demographic file but not in the enrollment file
- The number of procedures per patient seems low. Were all procedures pulled? (asked of 2 sites)
 - That's all the data we have
 - All procedures were pulled. Our recorded procedures are lower than other health plans, likely because the delivery system (pre-paid insurance plan) has no reimbursement incentive to record them
 - There is a new process for procedures. When providers order a procedure it remains open until completed and the order resulted. This requires the provider who performed the procedure to go into the electronic record and close it. The second step that is not being done. Therefore we are finding a huge number of orders that are never closed out, making it nearly impossible to tell if the procedure was done. Orders that are not closed out are not being counted.

Creating Study Datasets

Wrong Data Model

- Instead of specifying a simple and “raw” common data model, study asks sites to create highly-specified intermediate files
 - Sites implement intermediate files differently
 - Data checking was done manually; no distributed code for checking
- Many months of delays, no way to know if files are comparable

Study Datasets

Wrong Data Model: Too complicated

- **“Programming considerations.** To improve programming efficiency, all of the study periods may be identified first, with their corresponding t0s. Then, proceeding in date order, the control periods for that t0 are selected from persons who are not study drug users on t0 and are not already in the cohort. If the selected period comes from a window where there is future use of a study drug, then that period is removed from the set of study periods and the t0 appropriately reset. See section 7.2 Phase I Sample for more detail.”

More Examples

Outbreak of GBS

- Vaccine surveillance protocol required data owner to assist in development of study protocol and implement the complex logic
- Data owner created programming instructions to reflect local norms
- Highly summarized data sent to lead site for review
- Call CDC: ***13 cases of GBS among middle-aged men in NY who received a flu vaccine in October!***
- *Summary file populated with count of GBS claims, not unique people with a GBS claim*

More Examples

- Too many protocols, too many changes...version control
 - Protocol indicates chart review for all NDI identified deaths with a code of interest 30 days after t1, changed to 45 days after t1
 - Analyst opts to use 60 days “to be safe”
- Protocol lists tretinoin as an exclusion, with an asterisk... located two pages later clarifying: “Do not include isotretinoin (i.e., Accutane)”
 - Programmer misses asterisk, excludes patients using isotretinoin
 - Protocol doesn’t address forms of topical tretinoin: tretinoin microspheres and tretinoin emollient are inappropriately included on exclusion list at some sites
- Protocol asked sites to flag HPV vaccinees
 - Is a person considered ‘HPV vaccinated’ if they have received at least 1 vaccination in the series, or all three?

Multi-Institutional Research

Summary

- Possible to create comparable study datasets across institutions
- Many examples of successes in pharmacoepidemiology
- Success based on strong relationships and trust
- Always requires comprehensive data checking and expect revisions
- Need local expertise
- Seems like a waste to do it for a single project...

Multi-Institutional Research

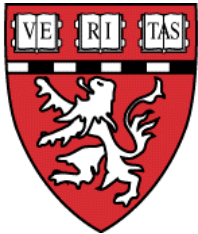
Recommendations

- Simple cohort definitions and data extraction
- Simple data structures that are re-usable
- Centralize study investigators and programmers
- Provide explicit instructions and code lists
- Detailed discussion of data model with analysts, variable by variable
- Coordinated analytics – programs distributed to sites for execution
- Tight control of study material
- Clear decision making and governance
- **A Re-usable Research Network**

Multi-Institutional Studies Using Observational Data: Opportunities and Challenges

Jeffrey Brown, PhD

The 32nd annual Midwest Biopharmaceutical Statistics Workshop
Ball State University, Muncie, IN
May 18 – 20, 2009



Department of Ambulatory Care and Prevention
Harvard Medical School and Harvard Pilgrim Health Care

