

How can PRO data from multiple EHRs be pooled?

Adam Wright, PhD



HARVARD
MEDICAL SCHOOL



BRIGHAM AND
WOMEN'S HOSPITAL

Dean F. Sittig, PhD



UTHealth[™]

The University of Texas
Health Science Center at Houston

**School of Biomedical
Informatics**

Potential Use Cases...

- **Multi-site, prospective, research protocol**
 - All subjects are enrolled in study
 - Mandatory study data are collected
 - PRO's sent to central coordinating center OR
 - PRO's analyzed at local site, aggregate data sent
- **Multi-site, retrospective, “big data” analysis**
 - Subjects are identified based on study criteria
 - All available data are collected for each subject
 - PRO's sent to central coordinating center OR
 - PRO's analyzed at local site, aggregate data sent

Pooling PRO data requires EHR Interoperability

- Data to be pooled are consistent
 - Source, method of collecting the data, patient population queried, timeframe, and questionnaires used
- Rigor of consistency depends on...
 - Specific aims
 - Composition and governance of the group
 - Clinical workflows used
 - Technical infrastructure and local info mgmt
 - Ethical and legal issues

Architectural Considerations for Pooling PRO Data

- **Centralized Data Warehouse**: sends all data extracted from EHRs into single, centralized data warehouse maintained by the data coordinating center
- **Distributed Data Warehouse**: data extracted from EHRs kept in a locally maintained data warehouse at each site.
 - Queries distributed to sites;
 - Summaries sent to the data coordinating center for analysis.

Centralized Data Warehouse

- Advantages:
 - Simpler technically, central quality control
 - Facilitates data analysis
 - Additional data analyses based on intermediate results
 - Reduces problems with small sample sizes

Centralized Data Warehouse

- Disadvantages:
 - Legal, regulatory, and proprietary issues of sharing data
 - Identification of duplicate patients
 - Reluctance to send patient-level data, unless fully de-identified

Distributed Data Warehouse

- Advantages:
 - Local organization retains control of data
 - No large centralized data warehouse
 - Sites can audit queries performed against their database

Distributed Data Warehouse

- Disadvantages:
 - Secure deduplication/overlay and record linkage is challenging
 - Ensuring data analyses are conducted accurately
 - Secondary data analyses based on intermediate findings difficult

Common Data Model and Reference Terminology

- Granularity of data?
- De-identification of data?
- Sequences, intervals, and episodes represented?
- Clinical domain(s) covered
- Governance of data model: updates?
- Standard interoperability reference terminologies supported?