NIH BD2K Think Tank
EHR Data Methodologies in Clinical Research: Perspectives from the Field

December 11, 2014
5635 Fishers Lane, Terrace Level Room 509-510
Rockville, MD

Draft Meeting Agenda

8:00 AM	**Welcome** NIH Co-organizers: Elaine Collier and Gina Wei
8:10 AM	**Charge for Meeting** Co-chairs: Michael Kahn and David Madigan
8:20 AM	**Session 1: Semantic Harmonization; Definition; Content; Ontologies** Panelists: George Hripcsak, Michael Kahn, Keith Marsolo, Marsha Raebel
10:10 AM	**Break**
10:20 AM	**Session 2: Multiple EHR for Single Participant; Other Multiple Data Sources** Panelists: Jason Doctor, Shawn Murphy, Nigam Shah, Jeffery Talbert
12:10 PM	**Lunch**
1:00 PM	**Session 3: Missing or Incomplete or Conflicting Data** Panelists: Miguel Hernan, David Madigan, Sally Morton, Alec Walker
2:50 PM	**Break**
3:00 PM	**Session 4: Longitudinal and other Temporal Issues for Long Term Studies** Panelists: Lesley Curtis, J. Michael Gaziano, Patrick Heagerty, Abel Kho
4:50 PM	**Wrap up**
5:30 PM	**Adjourn**

NIH BD2K Think Tank
EHR Data Methodologies in Clinical Research: Perspectives from the Field

December 11, 2014
5635 Fishers Lane, Terrace Level Room 509-510
Rockville, MD

Roster of Expert Panel Members

Lesley Curtis, PhD
Duke University

Jason Doctor, PhD
University of Southern California

J. Michael Gaziano, MD, MPH
Brigham and Women's Hospital

Patrick Heagerty, PhD
University of Washington

Miguel Hernan, MD, MPH, ScM, DrPH
Harvard School of Public Health

George Hripcsak, MD, MS
Columbia University

Michael Kahn, MD, PhD **(Co-chair)**
University of Colorado

Abel Kho, MD
Northwestern University

David Madigan, PhD **(Co-chair)**
Columbia University

Keith Marsolo, PhD
University of Cincinnati

Sally C. Morton, PhD
University of Pittsburgh

Shawn Murphy, MD, PhD
Massachusetts General Hospital

Marsha Raebel, PharmD
Kaiser Permanente Colorado

Nigam Shah, MBBS, PhD
Stanford University

Jeffery Talbert, PhD
University of Kentucky

Alec Walker, MD, DrPH
World Health Information Science Consultants

Co-Chairs: Michael Kahn, M.D., Ph.D. and David Madigan, Ph.D.
NIH Organizers: Elaine Collier, M.D. and Gina Wei, M.D., M.P.H.

NIH BD2K Think Tank
EHR Data Methodologies in Clinical Research: Perspectives from the Field

December 11, 2014
5635 Fishers Lane, Terrace Level Room 509-510
Rockville, MD

Meeting Purpose and Background

Purpose: This workshop will serve as a think tank to convene a small number of experts to specifically address methods for optimizing the robustness and use of data from the ***Electronic Health Records (EHR)*** for a variety of ***clinical research purposes*** that fall within NIH's domain. The think tank will recommend current strategies to address ***robustness*** and ***validity*** concerns or where new ***methodologies*** are needed to address these types of research studies. Traditional study design and statistical methods may need to be rethought in the context of using EHR data for research analysis. Where applicable, the think tank will take into account and build upon recommendations from the NIH workshop on "Enabling Research Use of Clinical Data".

Background: The expanded adoption of EHRs is generating growing volumes of data. This creates tremendous opportunity for enhancing the efficiency of conducting all types of clinical and population-based research. Nevertheless, clinical data from EHR differ from traditional forms of biomedical research data in important ways, including the fact that they were not collected for research purposes. Because of this, data robustness (including completeness, consistency) and validity for research "purpose" are major concerns. Additionally, the collection and use of data in these systems for research is complicated by policy issues as well as privacy and confidentiality concerns for patients and issues of public trust.

Focus: Given the potential broad scope of this field, this think tank will narrow its attention on:

- **EHR as the major data source**: The think tank will focus on EHR as the primary data source and not on other forms of clinical data such as mobile health technologies, smart devices, or high-throughput laboratory technologies and sensors that obtain data outside of current clinical settings. Where germane to the discussion of issues related to the use of EHR data for research purposes, it could be appropriate to discuss methodologies utilized in other types of data.

- **Dealing with imperfect EHR data as they are currently collected**: The think tank will focus on addressing issues on the 'back-end' of the EHR enterprise (i.e. the use of currently collected clinical data), rather than to develop strategies to improve the quality of clinical data on the 'front end" (e.g., accuracy of data entry by providers). Methods to improve the back end data by clarifying their metadata (i.e., time ordered vs time collected vs time assay performed) or to mine or understand free text documentation (e.g., via natural language processing) would be within the scope.

- **Wide range of clinical research studies**: The think tank will focus on methods around the robustness for "purpose" of current EHR data for a variety of clinical research studies, rather than focusing on any particular diseases. Study design methodology and statistical methods for data management and analysis are particularly relevant for this purpose. Types of clinical research domains could include: natural history observations; comparative effectiveness; genome-phenome; development and/or validation of risk factors, diagnostic tools, and clinical outcome assessments; etiology and mechanisms of diseases; complex multiple chronic conditions, co-morbidities, or treatment interactions; and studies in special populations.

- **Multi-site/Multi-system studies**: The think tank will focus on the use of data that can capture sufficiently large sample sizes of patient records. Methods that address issues for integration of EHR data with data in other systems are also particularly relevant.

NIH BD2K Think Tank
EHR Data Methodologies in Clinical Research: Perspectives from the Field

December 11, 2014
5635 Fishers Lane, Terrace Level Room 509-510
Rockville, MD

Charge to Each Session

Charge to Session 1: Semantic harmonization, definition, content, ontologies

Background: No two electronic health record data sources are identical, even when data are extracted from the same commercial vendor. Differences exist in how data are structured, stored, and represented. For example, one study lists 34 variations in the unit of measure for glycosylated hemoglobin (HbA1c) and 67 variations in the unit of measure for platelet count across just twelve data partners (Raebel, M et al Pharmacoepidemiol Drug Saf. 2014. 23 (6) 609-18)). Similar wide variations exist in EHR clinical documentation both within and across institutions. Vast differences in clinical workflows and institutional policies can have substantial impact how, when, and by whom data are recorded (e.g. structured versus unstructured; at triage versus during encounter; by medical assistant versus physician), all of which may alter the availiability, accuracy and interpretation of extracted data. Numerous methods have been developed to help reduce barriers to harmonizing disparate data across diverse data partners – some involve advanced technologies, others involve highly structured harmonization processes.

Purpose/Goal: The purpose of this panel is to describe "in-the-field" experiences and approaches to data harmonization that have been used in a wide range of successful EHR data sharing efforts. The goal is to ensure that meeting participants better understand the challenges of data harmonization in establishing meaningful and comparable data sharing networks.

Topics of potential interest:

- Methods for common understanding of content
- Automated methods for harmonization of disparate data
- Heterogeneous data and meta data handling
- Role of ontologies, terminologies, definitions
- Unstructured data context and imputing content meaning
- Other issues around special populations studies – pediatric, elderly, vulnerable, rare diseases, Tribal communities, mental health studies

Charge to Session 2: Multiple providers/EHRs for single participant; multiple other data sources

Background: The US health system is highly fragmented, resulting in patient-level data scattered across a wide range of clinical and administrative systems. For patients enrolled in clinical studies, additional regulatory and contractual barriers to research data present additional barriers to data access. Yet without access to complete data, critical interventions or outcomes can be missed leading to incorrect or incomplete inferences. As more patient data are available in electronic format, the **potential** exists to integrate and link clinical, billing, medication and specialized data from different data sources together.

Purpose/Goal: The purpose of this panel is to share experiences with combining data across multiple disparate data sources that have analyses to be performed that could not accomplished with just one data source. The goal is to ensure that meeting participants better understand the challenges and promises of data integration in establishing more complete patient- and population-level data.

Topics of potential interest:

- Multiple EHRs for data for participant needed concurrently, e.g., surgeon, primary care, specialist, rehabilitation, dialysis center
- Complex multiple diseases and conditions; co-morbidities
- Genome/Phenome
- EHR data and multiple other source data, e.g., genetic, environment, social, economic
- Other specialized clinical data – study and non-study specific
- Other issues around special populations studies – pediatric, elderly, vulnerable, rare diseases, Tribal communities, mental health studies

Charge to Session 3: Missing or Incomplete or Conflicting Data

Background: The scale of many of the more valuable clinical databases precludes any kind of manual curation so analysts must confront a variety of basic data challenges. Necessary data elements for many research studies are missing in many databases. Such missing elements include behavioral, psychosocial, and familial data that are not directly relevant to the clinical encounters, as well as various non-prescription medications, images, and/or laboratory values. Free text permeates the medical record, and natural language processing techniques remain somewhat primitive. Image analysis techniques are equally crucial. Conflicting data are common both within individual databases and across databases. Increasingly, patient records are scattered across multiple databases raising questions about privacy, linkage, and also conflicting data.

Purpose/Goal: The purpose of this panel is to share experiences with addressing missing, incomplete, or conflicting data that have been used in a wide range of successful EHR data analysis. The goal is to ensure that the meeting participants better understand the challenges of addessing missing, incomplete, or conflicting data in addressing various clinical and scientific questions.

Topics of potential interest:

- Improved methods needed – including natural language processing, as well as image analysis techniques
- Necessary data missing in common data models used in comparative effective research studies
- Missing, incomplete, or conflicting data within EHR data or between EHR and other study data
- Other issues around special populations studies – pediatric, elderly, vulnerable, rare diseases, Tribal communities, mental health studies

Charge to Session 4: Longitudinal and other Temporal Issues for Long-Term Studies

Background: Clinical databases generally provide time-stamped records. However, using EHR to generate robust research for longitudinal to uncover topics such as risk-factor identification, treatment effect, and disease prognosis is often fraught with challenges posed by the fragmented US health system and wide variations in care. The temporal nature of these data and the disparate sources of care raise methodological challenges for patient characterization, effect estimation, and predictive modeling. While significant progress has occurred in recent decades many basic issues remain unsolved. For example, methodology for causal inference from longitudinal data has progressed considerably, yet it has never been subjected to a high-fidelity evaluation and current computational approaches do not scale to large databases.

Purpose/Goal: The purpose of this panel is share experiences with leveraging datasets either within or across multiple data sources to address research questions that require longitudinal tracking of patients. The goal is to ensure that meeting participants better understand the challenges of addressing longitudinal and other temporal issues to enhance the use of EHR for long-term studies.

Topics of potential interest:

- Multiple screening protocols
- Incomplete follow-up
- Change in healthcare practices and policies over time
- Transfer of participants to new care system(s)/provider(s)
- Temporal relationships of data from multiple sources
- Retention and recruitment of participants
- Child transitions to adult provider