EHR Data Methodologies in Clinical Research: Perspectives from the Field

NIH Big Data to Knowledge (BD2K) Think Tank

Meeting Summary

BACKGROUND:
The broad adoption of Electronic Health Records (EHRs) is generating growing volumes of data. This data source creates tremendous opportunities for enhancing the efficiency of the conduct of a broad range of clinical and population-based research. Nevertheless, data from EHRs differ from traditional forms of biomedical research data in important ways, including the fact that they are not collected for research purposes. Thus, the robustness and validity of the data for research purposes raises concerns (including completeness and consistency). Additionally, the use of EHR data for research is complicated by policy issues as well as privacy and confidentiality concerns and issues of public trust.

PURPOSE AND SCOPE:
This think tank was convened on December 11, 2014 for researchers from the field to discuss issues around the methodologies for optimizing the robustness and use of Electronic Health Records (EHR) data for a variety of clinical research purposes.

Given the potential broad scope of this field, the discussion was intentionally focused on:

a) Data contained in EHRs, including relevant laboratory, testing, procedure, and medication data. Data obtained outside of the normal clinical setting, including those from mobile health technologies, smart devices, or high-throughput laboratory technologies and sensors, were not a focus. Integration of data from multiple EHRs, whether for an individual person or multiple individuals, is necessary to address many research questions. Integration of EHR data with data from other sources was also discussed.

b) The 'back-end' of the EHR enterprise (i.e. the use of currently collected clinical data), rather than developing 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 (e.g., time a laboratory test was ordered vs. time the biospecimen was collected vs. time assay was performed) or to mine or understand free text documentation (e.g., via natural language processing) were included in the discussion.

c) Methods and approaches for use of EHR data in the various types of clinical research of particular interest to NIH including, 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.

MEETING STRUCTURE:
Experts in accessing EHR data and experts in study design and analysis methods for research using EHR data presented the challenges, solutions, and needs related to four highlighted areas based on their experience and knowledge of the field. The highlighted sessions were: 1) semantic harmonization, definition, content, and ontologies; 2) multiple providers/EHRs for single participant; multiple other data sources; 3) missing or incomplete or conflicting data; and

4) longitudinal and other temporal issues for long-term studies. See the Agenda for additional details about the sessions. Each session included an extensive discussion period with the audience. In addition, the co-chairs moderated a wrap-up discussion with all attendees. The entire event was video cast live and is archived.

Based on the discussion, several major topic areas were identified as outlined below. Each area includes a number of recommendations and points to consider when using EHR data for clinical research purposes.

RECOMMENDATIONS AND POINTS TO CONSIDER:
The following list of recommendations and points to consider when using EHR as a data source for clinical research is not in any priority or temporal order.

STUDY DESIGN AND ANALYSES

- Selection of appropriate study design and analysis methods should be driven by the research question. Repurposing data collected for another purpose, including EHR data and data from other sources, requires careful consideration as to fitness-for-purpose of data and analysis method. This knowledge is necessary to select an appropriate study design and statistical analytic plan for valid interpretation of the final results. Even then, researchers should understand that statistical outputs such as confidence intervals and p-values may have operating characteristics that deviate from nominal levels. For instance, 95% confidence intervals may include true values substantially less (or more) than 95% of the time.
- Researchers using repurposed data should understand from those providing the data the assumptions, uncertainties, and limitations of the data and/or data model for their intended research purpose. Collaboration of researchers with data providers may reveal that the data or the data model utilized for its access is a poor fit for providing data to address the particular research question.
- A clear description of the rationale for use of the selected design and analysis plan including its advantages and disadvantages as well as the assumptions, uncertainties, and limitations of the data utilized should be part of every research plan and all manuscripts reporting the results. Every study should include extensive sensitivity analyses, demonstrating the robustness, or lack thereof, of the results to analytic design choices. Ideally this should include analyses of multiple data sources as well as exploration of different study designs.
- EHR data are intrinsically longitudinal yet many of the common analytical approaches ignore or insufficiently account for time. Research questions that use EHR data will become ever more complex and longitudinal methods for causal inference and prediction will play a central role. Adaptive study designs in longitudinal data also have considerable potential.
- Scalable and reproducible methods for identification of phenotypes in EHR data are much needed.

EHR AS A DATA SOURCE

- Electronic health records are constantly evolving with new functionality, leading to changes in data availability, data collection workflows, and database location.
- Harmonizing local data to a common data model is easy for some data domains, such as diagnoses and procedures, and difficult for other data domains, such as clinical observations and assessments.

- Large variations exist even in “coded” data, such as how a laboratory test unit of measure may be recorded. Changes in reference ranges and test names over time are common. Qualitative test results have no standardization. Prevalence of conditions, procedures, etc., can exhibit sharp and implausible changes over time.
- Extracting data from EHR into a common data model requires site-specific and vendor-specific (custom) programming and knowledge of local data collection practices/workflows.
- Defining study cohorts (phenotypes) is an iterative process, often requiring access to data elements found only in text notes (NLP: natural language processing).

DATA MODELS

- There is no perfect one-size-fits-all data model for addressing all research questions. Each data model has its strengths and weaknesses. The group emphasized the importance of using the data model that incorporates the necessary data and relationships that are most suitable for the research query. The raw data and all associated metadata irrespective of source should be preserved to allow incorporation into multiple data models.
- Although several data models are widely used, including common data models such as OMOP, i2b2 SHRINE, HMORN VDW, FDA Mini-Sentinel, and PCORnet,* these models will necessarily evolve and new models will emerge as novel data types and sources including wearable devices, intelligent home sensors, and social media are incorporated to address changes in clinical practice and ever more complex research queries. [*note: OMOP = Observational Medical Outcomes Project; i2b2 SHRINE = Informatics for Integrating Biology and the Bedside Shared Health Research Information Network; HMORN VDW = HMO Research Network Virtual Data Warehouse; PCORnet = Patient-Centered Clinical Research Network]
- New methods to facilitate intelligent algorithms that operate in real time for mapping data to various data models are needed to meet the increasing types of data and their interconnecting relationships to other data.
- A clear description of the rationale for the use of the selected data model including its advantages and disadvantages for addressing the research question should be part of every protocol and all manuscripts reporting the research results.

INTEGRATION OF DATA SOURCES

- Data from multiple sources are likely to be needed to address most research questions. The use of multiple data sources is encouraged around issues of missing, conflicting, and incomplete data from EHRs. Access to additional data on individuals such as billing, payer, pharmacy, registry, patient-provided, genomic, and public health data can provide more complete information to address the particular research question.
- Accurate approaches to matching individuals in multiple data sources while protecting privacy and confidentiality are needed to allow valid integration of data from multiple data sources.
- EHR data based studies targeting special populations, such as rare diseases, tribal communities, mental health studies, and pediatric or elderly populations, may prove especially challenging and require multiple data sources. Data fragmentation as individuals encounter different facets of the healthcare system represents an ongoing challenge.
- The combination of EHR and other observational data (e.g., billing, payer data) with data from randomized trials represents a critically important and largely unexplored frontier.

TRANSPARENCY

- The sharing of EHR-based phenotype algorithms enables researchers to leverage the extensive work already performed. However, phenotype algorithms may need to be modified when used for a purpose different from that for which they were originally designed. To enhance usability, phenotype algorithm documentation should be in readily computable formats as well as annotated text.
- Information about the quality of EHR data and which data model(s) is(are) utilized as well as study design and analytic methods should be readily available in the publication or report of research results utilizing EHR data. This permits researchers and readers to make an informed interpretation on the validity and generalizability of the results and conclusions, particularly if causal inferences are being made. Overwhelmingly, this information is not included in publications even when results are expected to impact clinical care.

COLLABORATION AND COMMUNICATION

- Collaboration and ongoing communication between those utilizing EHR data to address research questions, those providing the data, and those with expertise in study design and analysis are critical for research using EHR data. All bring different perspectives and expertise to address specific questions and therefore must be part of an integrated team needed for this type of research.
- Collaboration allows for most effective and efficient re-use of EHR data in addressing research questions. Additionally, the collaboration leads to refinement of data models and analysis methods and research questions most appropriate for use of EHR data.

PRIVACY

- Methods of data provision that optimize protection of privacy of individual level data in the data set should be utilized.
- Privacy cannot be an absolute guarantee. As new data sources are integrated, new threats to privacy may be enabled.
- Privacy protection is particularly challenging in studies of special populations, such as rare diseases, tribal communities, mental health studies, and pediatric or elderly populations. Sharing of effective models for use of EHR data in research that includes these populations is needed to inform practice.