NIH Collaboratory

Health Care Systems Research Collaboratory

Phenotypes, quality and data elements

Meredith Nahm, Rachel Richesson, and Ed Hammond

Rethinking Clinical Trials

Presentation

- Background and definitions
- Mission
- Related activities
- Interaction with demonstration projects
- Discussion

Phenotype definition

- An observable physical or biochemical characteristic of an organism as determined by both genetic makeup and environmental influence
- The expression of a specific trait, such as stature or blood type, based on genetic and environmental influences
  - *American Heritage Dictionary*
- The phenotype of an organism includes factors such as its physical appearance, the biochemical process that takes place in its body and its behavior as it lives in the world and interacts with other organisms.
- In short, the phenotype of an organism is the appearance it presents to observers.

In Lewis Carroll's *Through the Looking-Glass* Humpty Dumpty discusses semantics and pragmatics with Alice.

"I don't know what you mean by 'glory,' " Alice said.

Humpty Dumpty smiled contemptuously. "Of course you don't—till I tell you. I meant 'there's a nice knock-down argument for you!' "

"But 'glory' doesn't mean 'a nice knock-down argument'," Alice objected.

"When *I* use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean—neither more nor less."

"The question is," said Alice, "whether you *can* make words mean so many different things."

"The question is," said Humpty Dumpty, "which is to be master——that's all."

Alice was too much puzzled to say anything, so after a minute Humpty Dumpty began again. "They've a temper, some of them—particularly verbs, they're the proudest—adjectives you can do anything with, but not verbs—however, *I* can manage the whole lot! Impenetrability! That's what *I* say!"[

A working definition

- A phenotype represents a trait or characteristic that relates to the topic of interest.
- Synonyms include trait, characteristic, attribute, feature, artifact, or other. Phenotype seems to be in most common use.
- Word is important because of the potential of extracting knowledge from large data sets of electronic health records.

- Phenotypes may be viewed as a way of packaging knowledge in an understandable and usable way.

Phenotype Suite

- Phenotype signature
- Cohort identification
- Risk factor or disease precursor
- Trigger for test, treatment or other action
- Disease progression
- Outcome evaluation

Phenotype Signature

- The set of phenotypes that define a disease or conditions
- Expression may be a logic expression with AND, OR, NOT. Proposed similar to logic component of Arden Syntax, using XML.
- With research, components may be assigned weights
- Sum of weights produce a quantitative certainty factor
- Phenotypes define the data elements to be collected
- Weighting factors may vary as a consequence of data not available or logic component not satisfied.

Cohort Identification

- Phenotype set which identifies a group of patients as candidates for a particular controlled clinical research trial.
- Sets will be classified as the diagnostic codes identified by vocabulary source, clinical data, demographic data, environmental data, genomic data, other data, and constraints such as location.
- Diagnostic code sets will be hierarchically defined and will be defined from multiple controlled terminologies

Phenotypes sets for risk factors

- Phenotype set that defines the risk factors or disease precursors with attached waiting factors to the phenotypes
- Requires research to extract candidate phenotypes from EHRs may be a Big Data type project

Phenotype Trigger

- Phenotype set that defines conditions and events that may trigger a type of treatment, test or activity
- Probably not a new idea and concept is represented by many clinical decision support algorithm.

Disease Progression

- Phenotype set that defines indicators for potential progression of disease
- Requires research to extract candidate phenotypes from EHRs or verify/challenge
- May be a Big Data type project

Outcome Evaluation

- Phenotype set for defining the outcome measure for a particular course of treatment.

Table 1 Statistics

- Phenotype set to define the characteristics and statistics for a given data base.
- If Table 1 is to have comparative value, each site needs to use consistent definitions of disease, demographics, and other characteristics. Phenotype signatures would be reusable for these definitions.

Other considerations

- Create standard format for documenting phenotype sets
- Effectiveness will require the use of a common set of data elements with a rich set of data attributes , freely available
- Data elements should contain genomic, biomarkers, clinical, environmental, social, economic, geospatial, and any other kind of data that is involved in defining health and health care.
- Algorithms for expressing algorithms must be flexible and accommodate time and time intervals
- This approach permits and suggests follow up research projects to define and validate phenotype suites.
- Propose creating registries of phenotype suites

Related activity – eMERGE Network

- The Electronic Medical Records and Genomics (eMERGE) Network is a national consortium organized by NHGRI to develop, disseminate, and apply approaches to research.
- Combines DNA biorepositories with EHR systems for large-scale, high-throughput genetic research with the ultimate goal of returning genomic testing results to patients in a clinical care setting.
- Has published 13 algorithms for cohort identification
- Currently exploring more than a dozen new phenotype sets
- eMERGE focuses on ethical, legal, social, and policy issues such as privacy and interactions with the broader community.

eMERGE

- Core activities overlap but contains some different approaches
- Proposed creating a cooperative relationship with eMERGE Network
- Propose to work toward standards for defining phenotype sets including logic specification

Related Activity - Mini-Sentinel

- Mini-Sentinel is a pilot project sponsored by FDA to create an active surveillance system to monitor the safety of FDA-regulated medical products
- Mini-Sentinel uses a distributed data approach in which Data Partners retain control over data in their possession as a result of normal activities
- Developed and implemented the Mini-Sentinel Common Data Model (MSCDM) and the Mini-Sentinel Distributed Database (MSDD),

Mini-Sentinel

- Activities of core are related and complimentary
- Propose working with this group in developing common data model

The Landscape

- No standard data collection in EHRs
- What appears standard is not always so
  - Institutions have multiple sources of ICD-9-CM codes, lab values, and medication related data (e.g., orders vs. administration, clinical notes)
- The data reflect patient and clinician/organizational factors
- There is no standard template for phenotypes
- Common but not standard approaches
- They are iterative
- The concern is data quality and reproducibility

Phenotype Core - Charter

- The purpose of the core is to promote multi-disciplinary discussion and collaboration to advance the theory and practice for using Electronic Health Records (EHRs) to support clinical and health services research.
- Participants will share their experiences using EHR to support research in various disease domains and for various purposes. The HSC Phenotype Core group will identify and explore these experiences collectively to determine generalizable approaches, methods, and best practices (and suggest where tools are needed) to support the widespread use of consistent, practical, and useful methods to use widely available clinical data to advance health and healthcare research.
- The group will also explore and advocate for cultural and policy changes related to use of EHRs for clinical trials.

Members

- Includes representatives from interested Demonstration projects and the Collaborative Coordinating Center (Duke)
- Multi-disciplinary “experts”
  - Greg Simon – Suicide Study
  - Jennifer Robinson - Nighttime dosing study
  - Alan Bauck – Chronic Pain Study
  - John Dickerson– Chronic Pain Study
  - Chris Helker,– Hemodialysis Project
  - Denise Cifelli – TiME study
  - Rosemary Madigan – TiME Study
  - Jerry Sheehan – NIH
  - W. Ed Hammond - Duke Coordinating Center (DCC)
  - Meredith Nahm (DCC)
  - Rachel Richesson (DCC)

Interaction with Demonstration Projects

- Document use cases for phenotype sets for each demonstration project
  - Specify inclusion/exclusion criteria
- Standard definition of data elements to be used in the phenotype sets
- Explore literature for phenotype sets. Validate using EHRs.
- Review of other activities in the use and definition of phenotype sets
- Standard, model or template for definition of phenotype suites including phenotype signature which will be used to validate quality
- Understand and synthesize "Table 1's" for demonstration projects

Data Quality Work of the Core

1. Synthesize description of practices on demonstration projects
2. Report summary back to projects in context of relevant literature
3. Consult with demonstration projects on data quality assessment, data cleaning or enhancement

Initial look at Data Sources

External PRO	PRO in EHR	Research specific EHR screens	Data warehouse clinical data	Data warehouse admin data
Paper, interview	X		X	X
		X	X	X
			X	X
			X	X
	X		X	X
			X	X

One row per project

3 projects include patient or staff interviews

1 project is collecting data from an external lab

Initial Look at Data Quality Practices

Collection control	Completeness Ascertainment	Completeness % Complete	Accuracy Indiv. value	Accuracy Agg. indicator
Procedural Technical		Part of ETL into warehse	Part of ETL into warehse	
Procedural Technical		Yes, monthly monitoring	Health system validated	
	Indep. Data % Chart review		Call audit % Chart review	
		yes		Site2site variability
	% Chart review			
Procedural		yes	Valid values	

One row per project

Blank = not stated in initial application & text

Next steps

- Iteration of categorizations with demonstration projects
- Collection of more detailed information on data sources and data quality assessment practices
- More detailed characterization

Collaboratory and Core Impact

- Technical Challenges
  - Methods, tools, statistics, best practices
  - What is good enough?
- Culture changes
  - Are we prepared to accept what is good enough?
  - Practicality versus perfection

Discussion

Thank you.