NLM's current work in health data standards and interdependence between data structure, and semantic, standards

Clem McDonald, MD
Chief Health Data Standards Officer
National Library of Medicine
National Institutes of Health
U.S. Department of Health & Human Services

NCVHS Listening Session panel
August 25, 2021 | Webinar | 1:30-3:30pm EST

NIH U.S. National Library of Medicine

NIH/NLM'S CURRENT WORK

NLM's involvement with health data semantics

- LOINC - funding support
- SNOMED CT - funding support
- RxNorm/RxNav – develop and maintain coding system for medications including gerneric, brand names and classes (RxNav)
- MeSH - codes used primarily for indexing biomedical literature
- UMLS – A Metathesaurus of 200+ vocabulary/coding systems
- CDEs – common data elements for researchers
- VSAC – Value Set Authority Center; value sets initially required for CMS quality rules
- ClinVar – database of genomic variations and its relationship to human health
- dbSNP – database of human single nucleotide variations
- Taxonomy – database of classification and nomenclature for all organisms in public sequence databases (from E.coli to elephants all based on DNA/RNA)
- UCUM - units of measure code system– a web site for validating UCUM unit strings and converting values between commensurate units of measure

U.S. National Library of Medicine

NLMs support for health care data structures

- FHIR (Fast Healthcare Interoperability Resources) projects
  - Implementation of FHIR's Questionnaire/Structured Data Capture (SDC) (on-the fly web based input forms)
  - Implementation of FHIRpath, FHIR's Xpath-like underpinning for navigating and calculation
  - Research Data Finder - query tool for finding and pulling data from FHIR electronic medical records or research databases
  - Implementing the dbGaP database as FHIR resources. DbGap carries 5 B records with 210,000 different "names" but many are different names for the same concept. Very difficult with many structural dimensions embedded in unruly observation descriptions.

NIH has promoted Health Data Standards- a First

- 2018 Guide Notice encouraging researchers to explore the use of FHIR –Mostly structure
- 2020 Guide Notice encouraging the researchers to adopt and use the USCDI – the federal ONC specification of specific coding - mostly semantics (coding systems)
- Some funding for such efforts, as well

Underpinnings of NIH's interest in FHIR and other data standards

- NIH has had the goal of making research data interoperable since 2003
- This goal has been difficult to reach because availability of an accepted framework to standardize around was lacking
- FHIR has already been embraced by federal agencies and by the health insurance industry, EMR developers, and all the big health IT companies
- FHIR provides the needed framework and is perhaps the last and best great hope for NIH's ambition
- A committee of NIH funded investigators, NCPI (NIH Cloud Platform Interoperability), is already working towards this goal

SEMANTIC INTEROPERABILITY IS ONLY PART OF THE ANSWER

Some early developers thought semantics were everything

- They built systems that flowed from one word or phrase to another, stringing them together with codes that would convert to a narrative report
- At least five companies took this tactic
- Recording a note through these was impossibly slow for care providers who can dictate note at > 200 words per minute
- All five failed
- Weed's Problem-Oriented Medical Record promise was a very early version. It did work in obstetrics with a narrow scope. However, it failed miserably on the internal medicine ward because it took 6+ hours to write an admission note.
- Residents revolted, and U of Vermont deinstalled the system.

Semantics and structure are like love and marriage, horse and carriage

- As the song says, you can't have one without the other.
- They are quite dependent on one another
- Databases have structure made up of fields (or slots) –think of them as questions that take answers
- You can't decide the answers until you know the question and you often can't know what the question really means until you see the answers it wants
- If you don't anchor and standardize the structure, there is no hope for semantic interoperability
- (I know I am over simplifying)

The US has made progress on standardizing data structure

- V2, CDA and FHIR have defined the structures
- FHIR is the most elegant and advanced, but all three similar with respect to their tables/Resources
- Observations, for example, have a field for the Observation identifier, the relevant time of the observation, the abnormal flag, and the value in all of them
- We are closing in on a common structure
- Semantic operability occurs when everyone uses the same coding system (vocabulary)

More standardized semantics are also progressing

- USCDI has specified specific code systems (semantics) for specific fields.
  - LOINC to identify laboratory tests and other observation
  - SNOMED CT for problems, coded observation answers, and others
  - RxNorm for medications
  - UCUM for units of measure
- This round standardizes codes for a variety of clinical notes and reports
- Next round likely to standardize codes for many physical measures – 12-lead EKGS, tonometry, nerve conductions, endoscopies, and wall motion echo studies
- Lots of clean up work remains because idiosyncratic local code are still rife for most of these cases

The really tough remaining problem is codes and structure for H&P and progress notes

- This information comes from the minds of individual physicians, not for manufactured systems
- Entering this data would require scarce provider time
- So, let providers write their notes in free text and identify a very limited set of key variables (questions) for each problem, e.g., for heart failure : better/worse, body weight, and ejection fraction
- Depend on natural language processing (NLP) to pick off the anything else of interest from the free text

THE END