Data Engineering in Healthcare: progress and remaining challenges

Roberto A. Rocha, MD, PhD, FACMI
Managing Director
Semedy, Inc.
Assistant Professor of Medicine
Division of General Internal Medicine and Primary Care
Brigham and Women’s Hospital, Harvard Medical School

1st International Conference on AI in Healthcare (ICAIH): March, 2018
Disclosures

• Salary support from Semedy, Inc.
• Salary support from Wolters Kluwer Health (spouse)
Overview

- Motivation
- Background
- Interoperability
- Information models
- Conclusions
MOTIVATION
“Progress on the road toward integrating big data — both high-volume genomic findings and heterogeneous clinical observations — into practical clinical protocols and standard healthcare delivery requires that providers, HIT vendors, federated knowledge resources, and patients can ultimately depend upon those data being comparable and consistent.”

– “Absent comparability, the data are more or less by definition not able to support inferencing of any scalable kind ...”

– “Without consistency, users of complex biomedical data will have to spend added resources transforming the data into usable and predictable formats ...”

Challenges for AI

• “There is a **great deal of interest** in the potential of using the vast data sets represented in electronic health records (EHRs), in combination with AI algorithms …”
  – “… AI can perform with great accuracy when the relationship between diagnostic data and the diagnosis is well defined, when the relationship between the data and the diagnosis suffers from **error, variability or difficulty in discrimination**, AI algorithms also perform less well.”

• “**Extreme care is needed** in using EHRs as training sets for AI, where outputs may be **useless or misleading** if the training sets contain incorrect information or information with unexpected internal correlations.”

Relevance of data engineering

• “.. two viable solutions to address heterogeneous data:
  – defining a “common representation” and transforming all data into
    that common interlingua, or
  – adopting standards at the point of data generation to obviate the
    costs and confusion that often emerge from data transformation.”

• “... inferences will have hugely more power and accuracy if
  we aim big data methods at information that shares names
  and values”
  – “we do not want to waste analytic resources “discovering” that renal
    cancer behaves similarly to kidney cancer””

Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for
Promising preliminary results

• “Low-volume, structured clinical data contain sufficient information to train classifiers to perform near physician-level.”
  – 799 cases independently validated by more than 2 medical professionals (medical students and physicians)

• “Collecting such data is possible through human computation that is independently useful to clinicians.”

BACKGROUND
Data engineering?

• “The design, implementation, modeling, theory and application of database systems and their technology.” (IEEE TCDE)

• Purposeful design and implementation of models and related artifacts to ensure consistent, extensible, and interoperable data representation
  – Akin to knowledge and terminology engineering
Clinical data

• Highly complex
  – Diverse types of data
    ▪ structured and unstructured, images, sounds, etc.
  – Dynamic (changing) nature
    ▪ flexible and extensible underlying models

• Large quantities
  – High performance database environment
    ▪ response time is the critical factor

• Confidentiality and privacy (security)
Clinical systems (data) dichotomy

**EHRs/EMRs:** Electronic Health/Medical Record Systems

**PHRs:** Patient Health Records

**Medical Record Systems**

**Clinical Research Systems**

**EDCs:** Electronic Data Capture Systems

**eCRFs:** Electronic Case Report Forms
EHR systems

• Electronic Health Record (EHR)
  – “electronic version of a patients medical history, ... maintained by the provider over time, ... clinical data relevant to that persons care ... including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory ... radiology” (CMS definition)

• US market
  – Commercial EHRs dominate; small number of vendors
  – Designed for data storage & communication: human users
  – High tolerance for incomplete, incorrect, ambiguous data
  – Limited capability for computer-assisted decision making
  – Recent emphasis on data sharing (government incentives)
Types of clinical data

• Unstructured
  – “Mr. Jones has Appendicitis”
  – No structure or codes

• Structured
  – *Diagnosis*: “Appendicitis”
  – Question is defined (coded) but answer (value) is free-text

• Structured & Coded
  – *Diagnosis*: K35 *(Appendicitis)*
  – Question & answer are defined (coded)
Unstructured (narrative) data

- Significant portion of the medical record is available as narrative data (70-95%)
  - medical history, physical exam, progress notes, discharge summary, radiology reports, operative notes
- advantages: widespread, comprehensive, convenient, expressive, natural
- disadvantages: ambiguous, complex, different styles, redundant, embedded errors, loose structure
Clinical phenotyping data

• “Intrinsically complex, fraught with heterogeneity, and amply having the potential for enormous depth (many records).”

• “A single patient may have many thousands of unique attributes, each of which may have arbitrarily repeated measures.”

Data complexity

- Many rows
- Many columns
- Many rows & columns
- Few rows & columns
Why Data Engineering?

• Opportunity
  – Data defined with standard reference models – consistency, completeness, and interoperability
  – Sustainable process – new domains, single electronic record for all settings and disciplines

• Challenges
  – Data definitions not shared across EHR modules or settings – similar data stored and encoded differently
  – Large libraries of definitions – promote inconsistency, distinctions not documented, overlapping domains/topics
  – Manual verification – constantly evolving data collection tools (e.g. forms, flowsheets, templates, macros, etc.)
INTEROPERABILITY
Data & Knowledge interoperability

- Home
- Rehabilitation Phase
- Hospital Procedure
- Ambulatory Visit

Data + Knowledge flows between each phase.
Define “data points” (elements and values) using available standards

Define logical models that combine data points and provide meaningful clinical information

Obtain detailed provenance to understand how and when data was created, and also who provided the data

Represent semantic and temporal linkages between data instances
Clinical data: standards

- **Data elements and data values**
  - *Available*: reference terminologies and ontologies

- **Information models**
  - *Work in progress*: isolated efforts and collections
  - *Available*: clinical documents (multiple types)

- **Provenance models**
  - *Work in progress*: competing models

- **Semantic and temporal linkages**
  - *Preliminary work*

*Additional work is needed → opportunities*
Development of standards

• SDO: Standards Developing Organization
• Many national and international organizations
  – Interdependencies and overlapping efforts
• Several with specific focus on Healthcare
  – Examples: HL7, IHE, DICOM, CDISC, ...
LOINC

• Logical Observation Identifiers Names and Codes
• Organization: LOINC Committee
• Purpose: identification of laboratory and clinical observations (HL7 messages)
• Content: laboratory tests, clinical measurements, documents, etc.
• Information:
  – http://loinc.org
<table>
<thead>
<tr>
<th>LOINC</th>
<th>详称</th>
<th>成分</th>
<th>属性</th>
<th>时间</th>
<th>体系</th>
<th>精度</th>
</tr>
</thead>
<tbody>
<tr>
<td>24372-5</td>
<td>Peak systolic blood pressure --during right ventricular septal defect maximum velocity measurement</td>
<td>血管内脏收缩期高峰*在右心室间隔缺损最大速度测量过程中</td>
<td>压力或压强</td>
<td>时间点</td>
<td>动脉系统.XXX</td>
<td>定量型</td>
</tr>
<tr>
<td>75997-7</td>
<td>Systolic blood pressure by Continuous non-invasive monitoring</td>
<td>血管内脏收缩期</td>
<td>压力或压强</td>
<td>时间点</td>
<td>动脉系统</td>
<td>定量型</td>
</tr>
<tr>
<td>11378-7</td>
<td>Systolic blood pressure at First encounter</td>
<td>血管内脏收缩期</td>
<td>压力或压强</td>
<td>就医过程持续时间*第一</td>
<td>动脉系统</td>
<td>定量型</td>
</tr>
<tr>
<td>20185-5</td>
<td>End systolic blood pressure by US</td>
<td>血管内脏收缩期末期.*XXX</td>
<td>压力或压强</td>
<td>时间点</td>
<td>循环系统.XXX</td>
<td>定量型</td>
</tr>
<tr>
<td>20186-3</td>
<td>Peak systolic blood pressure by US</td>
<td>血管内脏收缩期高峰.*XXX</td>
<td>压力或压强</td>
<td>时间点</td>
<td>循环系统.XXX</td>
<td>定量型</td>
</tr>
<tr>
<td>24370-9</td>
<td>Peak systolic blood pressure --during mitral regurgitation maximum velocity measurement</td>
<td>血管内脏收缩期高峰*在二尖瓣反流最大速度测量过程中</td>
<td>压力或压强</td>
<td>时间点</td>
<td>动脉系统.XXX</td>
<td>定量型</td>
</tr>
<tr>
<td>24371-7</td>
<td>Peak systolic blood pressure --during aorta stenosis maximum velocity measurement</td>
<td>血管内脏收缩期高峰*在主动脉狭窄最大速度测量过程中</td>
<td>压力或压强</td>
<td>时间点</td>
<td>动脉系统.XXX</td>
<td>定量型</td>
</tr>
<tr>
<td>45372-0</td>
<td>Blood pressure systolic and diastolic--post phlebotomy</td>
<td>收缩期与舒张期血压*在静脉采血之后</td>
<td>压力或压强</td>
<td>时间点</td>
<td>动脉系统</td>
<td>定量型</td>
</tr>
<tr>
<td>50402-7</td>
<td>Blood pressure systolic and diastolic--after transfusion</td>
<td>收缩期与舒张期血压*在输血之后</td>
<td>压力或压强</td>
<td>时间点</td>
<td>动脉系统</td>
<td>定量型</td>
</tr>
<tr>
<td>50403-5</td>
<td>Blood pressure systolic and diastolic--before transfusion</td>
<td>收缩期与舒张期血压*在输血之前</td>
<td>压力或压强</td>
<td>时间点</td>
<td>动脉系统</td>
<td>定量型</td>
</tr>
</tbody>
</table>

Search generated 102 hits in 0.239 secs.

https://search.loinc.org
SNOMED CT

• Systematized Nomenclature of Medicine – Clinical Terms

• Organization: International Health Terminology Standards Development Organisation (IHTSDO)
  – SNOMED Terminology Solutions - College of American Pathologists

• Purpose: Encoding of multiple clinical domains

• Content: Comprehensive (diseases, signs, symptoms, living organisms, chemicals, body parts, morphology, occupations, modifiers, etc.)

• Information:
  – https://www.snomed.org
<table>
<thead>
<tr>
<th>Type at least 3 characters</th>
<th>Example: shou fra</th>
</tr>
</thead>
<tbody>
<tr>
<td>chest pain</td>
<td></td>
</tr>
</tbody>
</table>

79 matches found in 0.804 seconds.

**Chest pain**
- Dull chest pain
- Chest wall pain
- Acute chest pain
- Upper chest pain
- Chest pain rating
- Cardiac chest pain
- Central chest pain
- Chest pain at rest
- Ischemic chest pain
- Crushing chest pain
- Atypical chest pain

**Parents**
- Finding of region of thorax (finding)
- Pain of truncal structure (finding)

**Children (30)**
- Acute chest pain (finding)
- Atypical chest pain (finding)
- Cardiac chest pain (finding)
- Cardiac syndrome Y (finding)

http://browser.ihtsdotools.org
Many others (incomplete list)

- **RxNorm**: clinical drugs and drug delivery devices (NLM)

- **NDF-RT**: National Drug File - Reference Terminology (VA)

- **IIS**: Vaccination code sets (CDC)

- **HL7 Vocabulary domains** (messaging, documents, services)
Document standards

• Clinical Document Architecture (CDA)
• Organization: HL7 International
• Purpose: document markup standard that specifies the **structure** and **semantics** of "clinical documents" for the purpose of exchange between healthcare providers and patients
• Content:
  – Continuity of care, procedure note, patient assessments, etc.
  – Clinical oncology treatment plan, PHR plans, genetic testing reports, public cancer registries, etc.
  – Data Provenance
• Information:
Information models

• **Clinical Information Modeling Initiative (CIMI)**
• Organization: HL7 International
• Purpose: Improve the interoperability of healthcare systems through shared implementable clinical information models - a *single curated collection* with bindings to reference terminologies
• Content: laboratory test results, vital signs, diagnoses, procedures, patient measures, etc.
• Information:
  – [http://www.opencimi.org](http://www.opencimi.org)
INFORMATION MODELS
Standard data definitions:

Data Element (attribute): numeric measurement with unit

Birth Weight: <number><units>

LOINC
8339-4: Body weight^at birth
Mass; Pt; ^Patient; Qn; Measured

SNOMED CT (or UCUM)
258681007: Units of mass (SI)

Value set

SNOMED CT (or UCUM)
258682000: gram, g

Value (concept)

Standard data definitions: data elements + data values
### Coded data: variation

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem</td>
<td>Severe Pain</td>
</tr>
<tr>
<td>Pain</td>
<td>Severe</td>
</tr>
<tr>
<td>Severe Pain</td>
<td>Yes</td>
</tr>
<tr>
<td>Severe Pain</td>
<td>02-01-2001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finding</td>
<td>Elevated Sys BP</td>
</tr>
<tr>
<td>Sys BP</td>
<td>180 mmHg</td>
</tr>
<tr>
<td>Sys BP</td>
<td>Elevated</td>
</tr>
<tr>
<td>Sys BP</td>
<td>Abnormal</td>
</tr>
</tbody>
</table>
What information needs to be modeled?

- **All** clinical information within an EHR:
  - Allergies
  - Problems
  - Orders
  - Test results
  - Medication administration
  - **Physical exam** and **clinical measurements**
    - Signs, symptoms, diagnoses
  - Procedures
  - **Family history, medical history, and review of systems**
  - Clinical documents
Focus on relevant clinical topics

Figure 1. Criteria to Prioritize Clinical Topic Refinement

Acute Care Documentation (1/3)

• **Publication:**

• **Context:**
  – Large strategic initiative back in 2007 to develop **standardized acute care documentation** (*bedside*) across two major academic medical centers: Brigham and Women’s Hospital and Massachusetts General Hospital
• **Goals:**
  - Highly structured documentation to fulfill clinical needs, regulatory reporting, and data reuse
  - All clinical disciplines (e.g. nursing, medicine, social work, physical therapy, nutrition, occupational therapy)
  - Proactive data standardization in an effort to avoid ambiguity and duplication – e.g. naming convention for data elements, reuse of value sets, etc.

• **Results:**
  - Over 11,000 data elements defined, used in over 1,000 documentation templates – e.g. initial patient assessments, progress notes, procedure and perioperative notes, event notes, transfer notes, discharge notes, assessment scales, flowsheets, etc.
  - Bedside documentation system was not implemented
Acute Care Documentation (3/3)

- **Challenges:**
  - Clinical requirements *well understood* by stakeholder groups - easily gained traction when cited as a rationale for content development requirements
  - Data engineering requirements *not well understood* – formal processes to garner support and adherence
  - Limited resources, *expertise*, and competing priorities

- **Lessons learned:**
  - Assess *knowledge needs* and set *expectations* at the start of the project
  - Define an accountable *decision-making process*
  - Increase team *meeting moderation* skills
  - Ensure adequate *resources* and *competency training* with online collaborative tools
  - Develop *goal-oriented* teams and consultative *service-based* teams
Large-scale EHR implementation (1/4)

• **Publications:**


Large-scale EHR implementation (2/4)

• **Context:**
  – **System-wide** vendor EHR implementation (2012-2017) – replace existing clinical systems

• **Goals:**
  – Minimize *(resolve)* **inconsistent data definitions** across EHR applications and clinical settings, enabling and promoting **data reuse** and **interoperability**
  – **Practical** *(pragmatic)* approach to **governance** and **implementation** of structured data elements and reference models
    - Factors: resource allocation, implementation timeline, content refactoring, vendor best-practices, EHR limitations, etc.
Large-scale EHR implementation (3/4)

• Process:
  1. Identify clinical topics – align with strategic goals of the organization
  2. Create draft reference model(s) – find/consolidate/reuse models
  3. Quantify downstream data needs – reporting, regulatory requirements, clinical decision support, accurate billing, etc.
  4. Prioritize clinical topics – focus on high-value topics
  5. Validate reference model(s) – clinically accurate and complete
  6. Quantify gap with EHR content – prioritize revision/refactoring
  7. Disseminate validated model(s) – guide new content or revisions
  8. Request revisions to EHR content – change management process
  9. Assess reference model utilization – implementation and compliance
  10. Monitor for new evidence - revisions to reference model (evergreen)
Large-scale EHR implementation (4/4)

• **Results:**
  – Data elements: +**15,000** (forms) and +**45,000** (flowsheets)
  – Dedicated workgroup: +**5** reference models (*discontinued*)
    ▪ Pain Assessment: **47 data elements** organized into **9 data groups**
  – EHR system successfully implemented at all sites

• **Challenges:**
  – Implementation timeline **incompatible** with the development of detailed reference models
  – EHR **processes** and **tools** not designed to promote detailed, consistent, and reusable data definitions across applications and modules
  – EHR content & data refactoring is an **iterative** process that requires **expertise** and **motivated** individuals
CONCLUSIONS
Challenges (1/3)

• **Cost-effective semantic interoperability**
  – Existing standards make data exchange possible, but not simple or efficient (projects take *months* or *years*)
  – Data exchanged in a structured and coded format still represents a small portion of the electronic record
Challenges (2/3)

- Clinical systems that can seamlessly represent and process a **complete electronic patient care record**
  - Current systems frequently rely on legacy data architectures that limit the use of clinical models
  - Slow adoption of new technologies that can overcome the current data representation limitations
Challenges (3/3)

• Clinical models with proper **domain coverage** and **extensibility**
  – Existing methods and tools to use clinical models and ontologies are not accessible to typical clinicians
Opportunities

• Government providing **exceptional incentives** for Healthcare IT adoption
  – IT identified as a key enabler of a more effective healthcare system

• Proposed healthcare delivery models require high levels of **integration** within and across institutions
  – Moving towards **seamless collaboration** where patients are active contributors

• Opportunity for a **new generation of clinical systems** beyond efficient record storage and communication
  – New paradigm with **pervasive** computerized **data analysis** and **decision support**
  – Widespread use of interoperable services and data, with advanced functions that enable **team-based care**
Conclusions: implementation

• Early engagement of clinical leaders to set expectations of technical process, dependencies, and requirements

• Provision of formal training about informatics standards and governance processes

• Establish a data engineering team with proper authority and robust toolset – guide implementation and ensure compliance with processes and standards
Conclusions: data engineering

- Establish **governance** for essential clinical domains
- Seek alignment with **standards**, maximizing interoperability and external collaborations
- Define and optimize **processes** *(lifecycle)*
  - Implement software **platform** integrated with knowledge **sources** and **consumers**
- **Monitor** & **evaluate** processes and resulting models
- **Collaborate** with other institutions to help amortize operational **costs** and promote **innovation**
Participate!

• Understand the scope and applicability of existing standards
  – Gain access to available standards
  – Confirm how each standard applies to your organization

• Contribute to and influence the development of standards
  – Bring your specific needs and discuss implementation options
  – Seek information from other stakeholders to make informed decisions
  – Most SDOs welcome open and broad participation
    ▪ Healthcare providers, government stakeholders, payers, pharmaceutical companies, system vendors, consultants

• Achieve industry leadership by demonstrating interoperability
  – Learn about industry best practices
  – Understand implementation timeline and costs
  – Improve the quality and sustainability of your local systems
Acknowledgements

Saverio Maviglia
Sarah Collins
Karen Bavuso
Perry Mar
Li Zhou
Margarita Sordo
Members of the **Semedy** team

Beatriz Rocha
Stanley Huff
Thank you!

Roberto A. Rocha, MD, PhD
rarocha@bwh.harvard.edu
http://scholar.harvard.edu/rarocha

This work by Roberto A. Rocha is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License