Strategic Investments Towards Interoperability
Redwood MedNet, 25 July 2013

Christopher G. Chute, MD DrPH,
Professor, Biomedical Informatics, Mayo Clinic
Chair, ISO TC215 on Health Informatics
Chair, International Classification of Disease, WHO
Declarations

☐ No real or apparent financial conflicts of interest
  – All products are open-source

☐ Comments represent beliefs of the author
Secondary Use

Analyses or interpretation of clinical data across multiple patients

- Clinical Quality Improvement
- Comparative Effectiveness Analyses
- Outcomes Research
- Best Evidence Discovery
- Technology Assessment
- Data-driven Clinical Decision Support
From Practice-based Evidence to Evidence-based Practice

Data
- Patient Encounters
- Clinical Databases

Inference
- Registries et al.
- Medical Knowledge

Decision support
- Expert Systems
- Clinical Guidelines

“Secondary Use”
© 2012 Mayo Clinic

SHARPn.org
From Practice-based Evidence to Evidence-based Practice

Data

Clinical Databases

Registries et al.

Comparable and Consistent

Inference

Standards

Vocabularies & Terminologies

Medical Knowledge

Decision support

Patient Encounters

Expert Systems

Clinical Guidelines

Knowledge Management

“Secondary Use”

© 2012 Mayo Clinic
Comparable and Consistent Data

- Inferencing from data to information requires sorting information into categories
  - Statistical bins
  - Machine learning features

- Accurate and reproducible categorization depends upon semantic consistency

- Semantic consistency is the vocabulary problem
  - Almost always manifest as the “value set” problem
The Challenge

☐ Most clinical data in the United States is heterogeneous – non-standard
  – Within Institutions
  – Between Institutions

☐ Meaningful Use is mitigating, but has not yet “solved” the problem
  – Achieving standardization in Meaningful Use is sometimes minimized
U.S. Department of Health & Human Services
http://www.hhs.gov/

Office of the National Coordinator for Health Information Technology (ONC)
Program Official: Wil Yu http://healthit.hhs.gov

AREA 1
University of Illinois at Urbana-Champaign
(#10510624)
Security of Health IT
PI: Carl Gunter, PhD
http://sharps.org

AREA 2
The University of Texas Health Science Center at Houston
(#10510592)
Patient-Centered Cognitive Support
PI: Jiajie Zhang, PhD
http://sharpc.org

AREA 3
Harvard University (#10510924)
Healthcare Application and Network Platform Architectures
PI: Isaac Kohane, MD, PhD
Co-PI: Kenneth D. Mandl, MD, MPH

AREA 4
Mayo Clinic College of Medicine (#10510949)
Secondary Use of EHR Data
PI: Christopher Chute, MD, Dr. P.H
http://sharpn.org
SHARP Area 4: Secondary Use of EHR Data

- Agilex Technologies
- CDISC (Clinical Data Interchange Standards Consortium)
- Centerphase Solutions
- Deloitte
- Group Health, Seattle
- IBM Watson Research Labs
- University of Utah
- University of Pittsburgh
- Harvard Univ.
- Intermountain Healthcare
- Mayo Clinic
- Mirth Corporation, Inc.
- MIT
- MITRE Corp.
- Regenstrief Institute, Inc.
- SUNY
- University of Colorado
Cross-integrated, *open-source*, suite of projects and products

<table>
<thead>
<tr>
<th>Themes</th>
<th>Projects</th>
<th>Players</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Normalization</td>
<td>Clinical Data Normalization</td>
<td>IBM, Mayo, Utah, Agilex, Regenstrief</td>
</tr>
<tr>
<td>Phenotype Recognition</td>
<td>Natural Language Processing</td>
<td>Harvard, Group Health, IBM, Utah, Mayo, MIT, SUNY, i2b2, Pittsburgh, Colorado, MITRE</td>
</tr>
<tr>
<td>Data Quality &amp; Evaluation Frameworks</td>
<td>High-Throughput Phenotyping</td>
<td>CDISC, Centerphase, Mayo, Utah</td>
</tr>
<tr>
<td></td>
<td>UIMA &amp; Scaling Capacity</td>
<td>IBM, Mayo, Agilex, Mirth</td>
</tr>
<tr>
<td></td>
<td>Data Quality</td>
<td>Mayo, Utah</td>
</tr>
<tr>
<td></td>
<td>Evaluation Framework</td>
<td>Agilex, Mayo, Utah</td>
</tr>
</tbody>
</table>
SHARPn Tools

- Library or suite of open-source tools
  - Apache 2.0, Commercial friendly
- Positioned as Middle-ware
- Intended to work on EMR “messages”
  - Defined by Meaningful Use standards
  - HL7 V2.51, cCDA, CCD, etc.
  - Clinical text, narratives, reports
Modes of Normalization

- Generally true for both *structured* and *un-structured* data

- Syntactic transformation
  - Clean up message formats
  - HL7 V2, CCD/CDA, tabular data, etc
  - Emulate Regenstrief HOSS pipeline

- Semantic normalization
  - Typically vocabulary mapping
Transformation Target?

- Normalization begs a “normal form”
- Extant national and international standards do not fully specify
  - Focus on HIE or internal messaging
  - Canonical data representation wanting
  - Require fully machine manageable data
Clinical Data Normalization

- Data Normalization
  - Comparable and consistent data is foundational to secondary use

- Clinical Data Models – Clinical Element Models (CEMS)
  - Basis for retaining computable meaning when data is exchanged between heterogeneous computer systems.
  - Basis for shared computable meaning when clinical data is referenced in decision support logic.
**BloodPressurePanel**

**Description / Status:**

<table>
<thead>
<tr>
<th>Name</th>
<th>BloodPressurePanel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>BloodPressurePanel is an Associated CEM Panel that groups a systolic blood pressure, diastolic blood pressure, and mean arterial pressure all obtained at the same time.</td>
</tr>
<tr>
<td>Status</td>
<td>proposed</td>
</tr>
</tbody>
</table>

**RAW XML**

```xml
<ctype kind="panel" name="BloodPressurePanel" xmlns="" />
<key code="BloodPressurePanel_KEY_ECID" />
<item card="0-1" name="systolicBloodPressureMeasure" type="SystolicBloodPressureMeasure" />
<item card="0-1" name="diastolicBloodPressureMeasure" type="DiastolicBloodPressureMeasure" />
<item card="0-1" name="meanArterialPressureMeasure" type="MeanArterialPressureMeasure" />
<qual card="0-1" name="methodDevice" type="MethodDevice" />
<qual card="0-1" name="bodyLocationPrecoordin" type="BodyLocationPrecoordin" />
<qual card="0-1" name="relativeTemporalContext" type="RelativeTemporalContext" />
<qual card="0-M" name="patientPrecondition" type="PatientPrecondition" />
<mod card="0-1" name="subject" type="Subject" />
```
Data Element Harmonization

http://informatics.mayo.edu/CIMI/

- Stan Huff – CIMI
  - Clinical Information Model Initiative
- NHS Clinical Clinical Statement
- CEN TC251/OpenEHR Archetypes
- HL7 Templates
- ISO TC215 Detailed Clinical Models
- CDISC Common Clinical Elements
- Intermountain/GE CEMs
That Semantic Bit…

- Canonical semantics reduce to Value-set Binding to CEM objects
- Value-sets drawn from “standard” vocabularies
  - SNOMED-CT and ICD, LOINC, RxNorm
- Common Terminology Services (CTS2)
- NLM National Value-set Center
  - CTS2 Value-set services
Normalization Pipelines

- Input heterogeneous clinical data
  - HL7, cCDA/CCD, structured feeds

- Output Normalized CEMs
  - Create logical structures within UIMA CAS

- Serialize to a persistence layer
  - SQL, RDF, “PCAST like”, XML

- Robust Prototypes now posted
  - Early version production Q3 2012
SHARPN Data Normalization Architecture

Normalization pipeline:
1. Syntactic Parsing
2. HL7/CCD/CDA to Object form
3. Initializer
4. Map to UIMA types
5. CTAKEs (NLP)
6. Semantic Normalization
7. Generate CEM

Optional Component:
1. Healthcare system

Mirth Connect Interface Engine:
1. Data transmitted via NwHIN
6. Document DB CEM DB

EMPI:
1. Batch CEM Processing
2. EMPI Patient Correlation
NLP Deliverables and Tools

http://informatics.mayo.edu/sharp/index.php/Tools

- cTAKES Releases
  - Smoking Status Classifier
  - Medication Annotator
- cTAKES Side Effects module
- Modules for relation extraction

- Integrated cTAKES (icTAKES)
  - an effort to improve the usability of cTAKES for end users

- NLP evaluation workbench
  - the dissemination of an NLP algorithm requires performance benchmarking. The evaluation workbench allows NLP investigators and developers to compare and evaluate various NLP algorithms.

- SHARPn NLP Common Type
  - SHARPn NLP Common Type System is an effort for defining common NLP types used in SHARPn; UIMA framework.

© 2012 Mayo Clinic
Specialized NLP Tools

- MedTagger – contextual clinical concepts extraction (certainty, status, and experiencer) – (i2b2 winner)
- MedXN – medication detection and normalization tool that associates medication mentions to clinical drugs in RxNorm
- SharpnDN – normalization clinical information across diverse data sources to common format
- MedTime – detect and normalize time expression from clinical narrative- The best system internationally
- meTAKES – an integrated NLP front-end for accessing data
  - Semantic, Indexing, Just-in-time-annotation, multi-layer language information retrieval
- Many phenotype-specific NLP tools – DILI, HF, PAD, DVT, Asthma

© 2012 Mayo Clinic
NLP Common Type Systems
High-Throughput Phenotyping

- Phenotype - a set of patient characteristics:
  - Diagnoses, Procedures
  - Demographics
  - Lab Values, Medications

- Phenotyping – overload of terms:
  - Originally for research cohorts from EMRs
  - Obvious extension to clinical trial eligibility
  - Quality metric Numerators and denominators
  - Clinical decision support - Trigger criteria

© 2012 Mayo Clinic
Cost per Raw Megabase of DNA Sequence

Moore's Law

National Human Genome Research Institute

[genome.gov/sequencingcosts](http://genome.gov/sequencingcosts)
EMR Phenotype Algorithms I

- Typical components
  - Billing and diagnoses codes; Procedure codes
  - Labs; Medications
  - Phenotype-specific co-variates (e.g., Demographics, Vitals, Smoking Status, CASI scores)
  - Pathology; Imaging?

- Organized into inclusion and exclusion criteria

- Experience from eMERGE Electronic Medical Records and Genomics Network (http://www.gwas.net)
EHR-driven Phenotyping Algorithms II

1. Diabetes diagnosis (T1 or T2)
   - Yes
   - No
      - Exclude

2. DR/ME in Diagnoses or Problem Lists
   - DR/ME ICD9 Code
     - Yes
     - No
       - Exclude
   - DR/ME Problem List
     - No
     - Exclude

3. Negative Mention of DR/ME
   - Yes
   - No
     - Case

4. Eye exam within past 2 years
   - Yes
   - Control

Phenotype Algorithm

Rules

Evaluation

Visualization

Transform

Transform

Data

Mappings

NLP, SQL

[eMERGE Network]

© 2012 Mayo Clinic
Modeling and Executing Electronic Health Records Driven Phenotyping Algorithms using the NQF Quality Data Model and JBoss® Drools Engine

Dingcheng Li, PhD¹, Gopu Shrestha, MS¹, Sahana Murthy, MS¹ Davide Sottara, PhD² Stanley M. Huff, MD³ Christopher G. Chute, MD, DrPH¹ Jyotishman Pathak, PhD¹
¹Mayo Clinic, Rochester, MN  ²University of Bologna, Italy  ³Intermountain Healthcare, Salt Lake City, UT

Abstract

With increasing adoption of electronic health records (EHRs), the need for formal representations for EHR-driven phenotyping algorithms has been recognized for some time. The recently proposed Quality Data Model from the National Quality Forum (NQF) provides an information model and a grammar that is intended to represent data collected during routine clinical care in EHRs as well as the basic logic required to represent the algorithmic criteria for phenotype definitions. The QDM is further aligned with Meaningful Use standards to ensure that the clinical data and algorithmic criteria are represented in a consistent, unambiguous and reproducible manner.

However, phenotype definitions represented in QDM, while structured, cannot be executed readily on existing EHRs. Rather, human interpretation, and subsequent implementation is a required step for this process. To address this need, the current study investigates open-source JBoss® Drools rules engine for automatic translation of QDM criteria into rules for execution over EHR data. In particular, using Apache Foundation’s Unstructured Information Management Architecture (UIMA) platform, we developed a translator tool for converting QDM defined phenotyping algorithm criteria into executable Drools rules scripts, and demonstrate their execution on real patient data from Mayo Clinic to identify cases for Coronary Artery Disease and Diabetes. To the best of our knowledge, this is the first study illustrating a framework and an approach for executing phenotyping criteria modeled in QDM using the Drools business rules management system.

Introduction

Identification of patient cohorts for conducting clinical and research studies has always been a major bottleneck and time-consuming process. Several studies as well as reports from the FDA have highlighted the issues in delays with subject recruitment and its impact on clinical research and public health. To meet this important requirement, increasing attention is being paid recently to leverage electronic health record (EHR) data for cohort identification. In particular, with the increasing adoption of EHRs for routine clinical care within the U.S. due to Meaningful Use, evaluating the strengths and limitations for secondary use of EHR data has important implications for clinical and translational research, including clinical trials, observational cohorts, outcomes research, and comparative effectiveness research.

In the recent past, several large-scale national and international projects, including eMERGE, SHARPn and i2b2, are developing tools and technologies for identifying patient cohorts using EHRs. A key component in this process is to define the “pseudocode” in terms of subject inclusion and exclusion criteria comprising primarily semi-structured data fields in the EHRs (e.g., billing and diagnoses codes, procedure codes, laboratory results, and medications). These pseudocodes, commonly referred to as phenotyping algorithms, also comprise logical operators, and are in general, represented in Microsoft Word and PDF documents as unstructured text. While the algorithm development is a team effort, which includes clinicians, domain experts, and informaticians, informatics and IT experts often operationalize their implementation. Human intermediary between the algorithm and the EHR system is required for primarily two main reasons: first, due to the lack of formal representation or specification language used for modeling the phenotyping algorithms, a domain expert has to interpret the algorithmic criteria. And second, due to the disconnection between how the criteria might be specified and how the clinical data is represented within an institution’s EHR (e.g., algorithm criteria might use LOINC® codes for lab measurements, whereas the EHR data might be represented using a local or proprietary lab code system), a human interpretation is required to transform the algorithm criteria and logic into a set of queries that can be executed on the EHR data.

To address both these challenges, the current study investigates the Quality Data Model proposed by the National Quality Forum along with open-source JBoss® Drools rules management system for the modeling and execution of EHR-driven phenotyping algorithms automatically. QDM is an information model and a grammar that is intended to represent data collected during routine clinical care in EHRs as well as the basic logic required to...
Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project

Susan Rea, Jyotishman Pathak, Guergana Savova, Thomas A. Oniki, Les Westberg, Calvin E. Beebe, Cui Tao, Craig G. Parker, Peter J. Haug, Stanley M. Huff, Christopher G. Chute
SHARP and Beacon Synergies

- SHARP will facilitate the mapping of comparable and consistent data into information and knowledge.

- SE MN Beacon will facilitate the population-based generation of best evidence and new knowledge.

- SE MN Beacon will allow the application of Health Information Technology to primary care practice:
  - Informing practice with population-based data
  - Supporting practice with knowledge
Principles and Conclusions

- Comparable and Consistent Data
- Importance of Canonical Form
  - At source or by transformation
- Standards are Evolving – Rapidly
- Meaningful Use is Good Start
  - Continued evolution expected
- SHARPn tooling facilitates
  - Normalization/NLP, Phenotyping

SHARPn.org

© 2012 Mayo Clinic