Data Harmonization: Challenge, Options, Strategy

Digital Data Priorities for Continuous Learning in Health and Health Care

IOM Workshop, March 23, 2012

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

SHARPN.org
Strategic Health IT Advanced Research Projects (SHARP) Program
From Practice-based Evidence to Evidence-based Practice

Data

Clinical Databases

Registries et al.

Inference

Comparable and Consistent Vocabularies & Terminologies

Standards

Medical Knowledge

Decision support

Patient Encounters

Expert Systems

Clinical Guidelines

Knowledge Management

"Secondary Use"

SHARPn.org

© 2012 Mayo Clinic
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://sharpm.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 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>
Mission

To enable the use of EHR data for secondary purposes, such as clinical research and public health.

Leverage health informatics to:
- generate new knowledge
- improve care
- address population needs

To support the community of EHR data consumers by developing:
- open-source tools
- services
- scalable software
Normalization Options

- Normalize data at source
  - Fiat, regulation, patient expectations

- Transformation and mapping
  - Soul of “ETL” in data warehousing

- Hybrid (graceful evolution to source)
  - “New” systems normalize at source
  - Transformation of legacy system data
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

Dr. Huff on Data Normalization

Stanley M. Huff, M.D.; SHARPn Co-Principal Investigator; Professor (Clinical) - Biomedical Informatics at University of Utah - College of Medicine and Chief Medical Informatics Officer Intermountain Healthcare. Dr. Huff discusses the need to provide patient care at the lowest cost with advanced decision support requires structured and coded 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.
A diagram of a simple clinical model

Clinical Element Model for Systolic Blood Pressure

- **SystolicBP**
  - data: 138 mmHg
  - qals
  - **BodyLocation**
    - data: Right Arm
  - **PatientPosition**
    - data: Sitting
BloodPressurePanel

**Description / Status:**

- **Name:** BloodPressurePanel
- **Definition:** 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.
- **Status:** proposed

**Details / XML View:**

```xml
<cttype kind="panel" name="BloodPressurePanel" xmlns=""
<key code="BloodPressurePanel_KEY_ECID" />
<item card="0-1" name="systolicBloodPressureMeas" type="SystolicBloodPressureMeas" />
<item card="0-1" name="diastolicBloodPressureMeas" type="DiastolicBloodPressureMeas" />
<item card="0-1" name="meanArterialPressureMeas" type="MeanArterialPressureMeas" />
<qual card="0-1" name="methodDevice" type="MethodDevice" />
<qual card="0-1" name="bodyLocationPrecoord" type="BodyLocationPrecoord" />
<qual card="0-1" name="bodyPosition" type="BodyPosition" />
<qual card="0-M" name="relativeTemporalContext" type="RelativeTemporalContext" />
<qual card="0-M" name="patientPrecondition" type="PatientPrecondition" />
<mod card="0-1" name="subjct" type="Subjct" />
```
Data Element Harmonization

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

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

- Recognize that use-case specific work-flow enters into CEM-like objects
  - Clinical EHR implementation
  - CLIA or FDA regulatory overhead

- Secondary Use tends to focus on data

- Create “core” CEMs
  - Labs, Rxs, Dxs, Pxs, demographics
That Semantic Bit…

- Canonical semantics reduce to Value-set Binding to CEM objects
- Value-sets should obviously be drawn from “standard” vocabularies
  - SNOMED-CT and ICD
  - LOINC
  - RxNorm
- But others required: HUGO, GO, HL7
Value-sets: Few or Many

- Everybody wants “manageable” number of value-sets
  - Estimates of 6-12
- Likely we will require thousands
- Raises requirement for terminology services and national repository
  - Once and future “US Realm”
On Mapping

- Semantic mapping from local codes to “standard” codes is required
  - Not magic, humanly curated
- Reality of idiosyncratic local codes is perverse
  - Why does every clinical lab in the country insist on making up its own lab codes?
Normalization Pipelines

- Input heterogeneous clinical data
  - HL7, CDA/CCD, structured feeds
- Output Normalized CEMs
  - Create logical structures within UIMA CAS
- Serialize to a persistence layer
  - SQL, RDF, “PCAST like”, XML
- Robust Prototypes exist
  - Early version production Q3 2012
This slide is obvious

1. Physical Quantity
2. Coded Values
3. Text field

Determine payload
Is laboratory data
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.
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
EMR Phenotype Algorithms

- 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)
Drools-based Architecture

- Clinical Element Database
- Data Access Layer
- Transformation Layer
- Business Logic
  - Inference Engine (Drools)
  - Service for Creating Output (File, Database, etc)
- List of Diabetic Patients

Transform physical representation → Normalized logical representation (Fact Model)

Jyotishman Pathak, PhD, discusses leveraging open-source tools such as Drools.
Phenotyping Activities

- **DROOLS**
  - Prioritize “Drools-izing” eMERGE algorithms (Diabetes, PAD and Hypothyroidism) and PGRN algorithms
  - Role of Drools for implementing the quality measures

- **Phenotyping Workbench / PhenoPortal**
  - develop an implementation independent, phenotyping logic representation template for algorithm design
  - Role of CEMs and NQF Quality Data Model (QDM)
  - Publicly accessible Web-based library for phenotyping algorithms
  - Phenotyping Graphical User Interface or “plug & play” workbench for algorithm design and evaluation

- **Just-In Time Phenotyping**
  - Apply algorithms as “data sniffers” that can be plugged within an UIMA pipeline
  - Online, real-time phenotyping (e.g., for clinical decision support)

- **Machine Learning Phenotyping**
  - leverage machine learning methods for rule/algorithm development, and validate against expert developed ones
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
More Information

SHARP Area 4: (SHARPn) Secondary Use of EHR Data

www.sharpon.org

Southeast Minnesota Beacon Community

www.semnbeacon.org