Data Normalization

Hongfang Liu
Data-driven Healthcare

Big Data

Analytics

Domain Pragmatics

Experts

Knowledge

Research

Practice
A framework for secondary use of EMRs

Production Systems

Production Databases

Replicate

Replicate

Query

Enterprise Repository/Data Warehouse

Data Normalization

Query

Workflow or goal specific

Workgroup Datamarts

Query

Query

Query
SHARPn DN Goals

• To conduct the science for realizing semantic interoperability and integration of diverse data sources
• To develop interoperable tools and resources enabling the generation of normalized EMR data for secondary uses
Data Normalization

Normalization Targets

Normalization Process

Information Models

Target Value Sets

Tooling

Raw EMR Data

Normalized EMR Data
Normalization Targets

- Clinical Element Models
  - Based on Intermountain Healthcare/GE Healthcare’s detailed clinical models

- Terminology/value sets associated with the models
  - Using standards where possible
Normalization Process

- Configuration of Model (Syntactic) and Terminology (Semantic) Mapping
- UIMA Pipeline to transform raw EMR data to normalized EMR data based on mappings
Four Subprojects

• Clinical Information Modeling
• Value Sets Management (CTS2)
• End-to-End Pipeline
• Normalized Data Representation and Store
Four Subprojects

• Clinical Information Modeling
• Value Sets Management (CTS2)
• End-to-End Pipeline
• Normalized Data Representation and Store
Embracing the fact that data may not be able to be normalized and enabling bottom-up and top-down
Status of Secondary Use CEMs

- Model specification is final
- CEM Browser is in production
- Manuscript is in preparation

Future:
Secondary Use CEMs and CEM Browser will be maintained through Clinical Information Modeling Initiative (CIMI)
SecondaryUseNotedDrug – Output (1/2)
SecondaryUseNotedDrug – Output (2/2)

```xml
<formulation usage="qualifier">
  <CD>
    <code>
      <value>10311</value>
    </code>
    <codeSystem>
      <value>2.16.840.1.113883.6.88</value>
    </codeSystem>
    <codeSystemVersion>Feb 2nd 2012 RXNORM</codeSystemVersion>
    <originalText>TAB</originalText>
  </CD>
</formulation>
```
NLP in data normalization

• A large amount of clinical information is in clinical narratives, NLP is a critical component in data normalization

• cTAKES has been wrapped into the data normalization pipeline to normalize data in clinical narratives
The release of the end-to-end DN framework is planned on October 1, 2013.
CEM Representation and Store

- Natively XML document
- Relational database or document database?
- Release 2.0 will be in couchDB
- Implemented: RDF and JSON
- Planned: rendering CCDA documents from CEM instances through OHT tooling
In summary

• Secondary use data model
• CTS2
• Normalization framework
• Data representation and store