Building a Robust, Scalable and Standards-Driven Infrastructure for Secondary Use of EHR Data: The SHARPn Project

Susan Rea[a,*], Jyotishman Pathak[b], Guergana Savova[c], Thomas A. Oniki[d], Les Westberg[e], Calvin E. Beebe[b], Cui Tao[b], Craig G. Parker[a], Peter J. Haug[a,f], Stanley M. Huff[d,f], Christopher G. Chute[b]

[a]Homer Warner Center for Informatics Research, Intermountain Healthcare, Murray, UT, USA
[b]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[c]Childrens Hospital Boston and Harvard Medical School, Boston, MA, USA
[d]Intermountain Medical Center, Intermountain Healthcare, Murray, UT, USA
[e]Agilex Technologies, Inc., Chantilly, VA, USA
[f]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

* Corresponding author: Address: Homer Warner Center for Informatics Research, Intermountain Medical Center, 5171 Cottonwood St,, Murray, UT 84107. 801-442-0748 (fax). Email address: Susan.Rea@imail.org.

Keywords
electronic health records
meaningful use
health information exchange
high-throughput phenotyping
natural language processing

ABSTRACT

The Strategic Health IT Advanced Research Projects (SHARP) Program, established by the Office of the National Coordinator for Health Information Technology in 2010 supports research findings that remove barriers for increased adoption of health IT. The improvements envisioned by the SHARP Area 4 Consortium (SHARPn) will enable the use of the electronic health record (EHR) for secondary purposes, such as care process and outcomes improvement, biomedical research and epidemiologic monitoring of the nation's health. One of the primary informatics problem areas in this endeavor is the standardization of disparate health data from the nation's many health care organizations and providers. The SHARPn team is developing open source services and components to support the ubiquitous exchange, sharing and reuse or 'liquidity' of operational clinical data stored in electronic health records. One year into the design and development of the SHARPn framework, we demonstrated end to end data flow and a prototype SHARPn platform, using thousands of patient electronic records sourced from two large healthcare organizations: Mayo Clinic and Intermountain Healthcare. The platform was deployed to (1) receive source EHR data in several formats, (2) generate structured data from EHR narrative text, and (3) normalize the EHR data using common detailed clinical models and Consolidated Health Informatics standard terminologies, which were (4) accessed by a phenotyping service using normalized data specifications. The architecture of this prototype SHARPn platform is presented. The EHR data throughput demonstration showed success in normalizing native EHR data, both structured and narrative, from two independent organizations and EHR systems. Based on the demonstration, observed challenges for standardization of EHR data for interoperable secondary use are discussed.

1. INTRODUCTION

Strategic Health IT Advanced Research Projects (SHARP) were established under the direction of the Office of the National Coordinator for Health Information Technology (ONC).[1, 2] SHARP research and development is focused on achieving breakthrough advances to address well-documented problems that have impeded adoption of the electronic health record (EHR):

- Security and privacy of health information
- User interfaces that support clinical reasoning and decision-making
- Shared application and network architectures
- Secondary use of EHR data to improve health [3, 4]

The health improvements envisioned by the secondary-use SHARP consortium (SHARPn) include care process and outcomes measurement, more efficient biomedical research and support for public health activities. Critical to these secondary-use scenarios is the transformation of health information into standards-conforming, comparable, and consistent data. Traditionally, a patient's medical information, such as medical history, exam data, hospital visits and physician notes, are stored inconsistently and in multiple locations. SHARPn will enable the use of EHR data for secondary purposes by creating tangible, scalable, and open-source tools, services and software for large-scale health record data normalization and sharing.[5]

One year into the design and development of the SHARPn framework, an integrated set of services and components was implemented as a prototype platform for natural language processing (NLP) and normalization pipelines to a persisted data store to cultivate high-throughput phenotyping. A demonstration of the prototype platform, focused on the throughput of normalized data, was conducted. Actual data from approximately ten thousand patient records were input from two independent EHRs: the Mayo Clinic and Intermountain Healthcare. EHR data corresponded to input data for a set of diabetes phenotyping rules used in practice. EHR data were sourced from both structured data and provider clinical narratives. Data were normalized to consistent, standardized forms in the SHARPn NLP and normalization components and stored in a shared database. A rules engine queried cases in the database for input data defined for a diabetes phenotyping algorithm. The target for this data throughput demonstration was the successful population of standardized data in the phenotyping data model.

In this paper, we present an overview of the architecture of this prototype SHARPn platform, describe the data flow from EHR to standardized data inputs for phenotyping, and discuss the challenges for shared, standardized secondary EHR data usage as observed in this data throughput demonstration.

2. BACKGROUND: THE SHARPn FRAMEWORK

2.1. Data Normalization and Phenotyping Architecture

The prototype platform supports a process flow implemented for this demonstration project. The SHARPn architecture consists of an adaptable framework of components and services. An overview of the SHARPn prototype data normalization and phenotyping architecture is shown below:

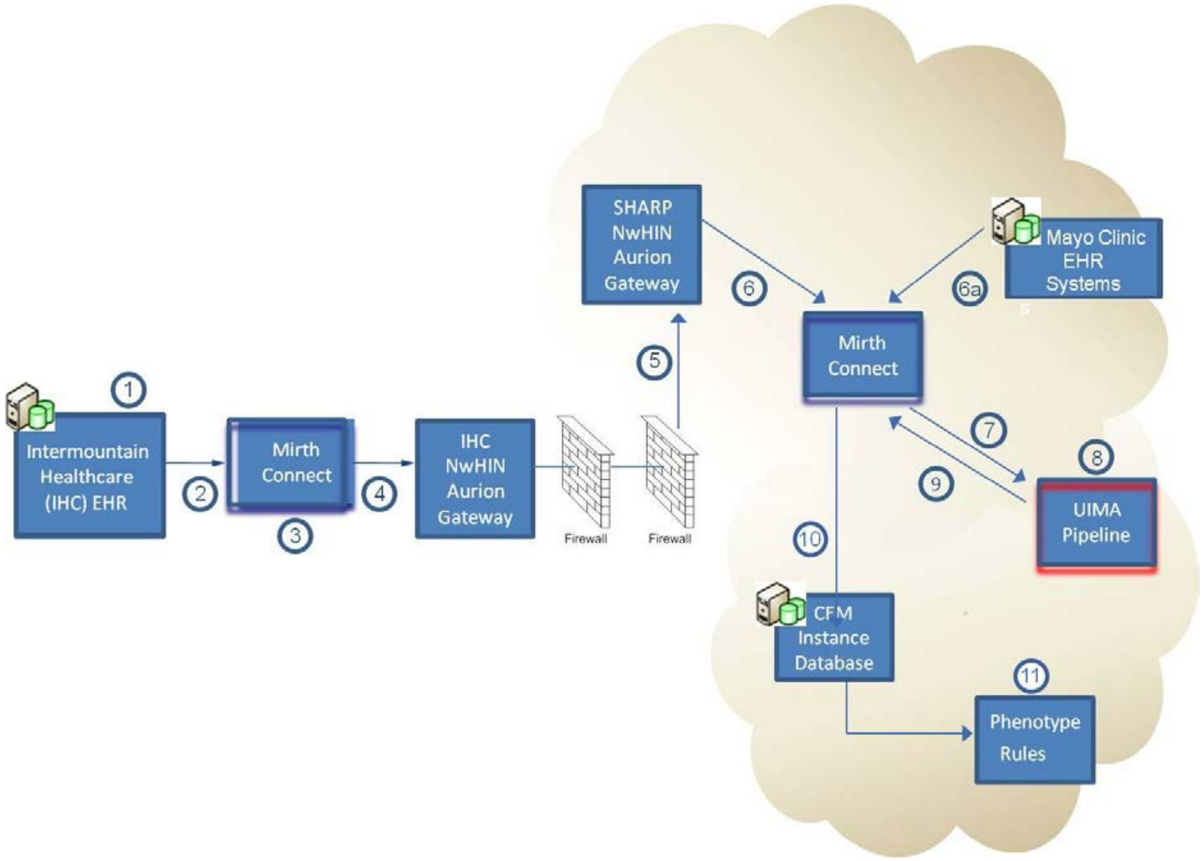

Figure 1. SHARPn Data Normalization and Phenotyping Architecture

An overview of all processes and data flows (Figure 1) is given in brief:

1. Intermountain Healthcare (IHC) structured clinical data were generated in Health Level Seven (HL7)[6] 2.x format. Other data input formats are supported in the architecture. Encounter ICD-9-CM[7] codes and attributions were prepared in tabular text files.
2. IHC data were pushed to Mirth Connect[8] on the IHC side of its firewall.
3. Mirth Connect is an open source product created by Mirth Corporation to transform messages among formats and/or route them from one location to another. Mirth Connect creates Nationwide Health Information Network (NwHIN)[9] Document Submission (XDR) messages from HL7 2.x and ICD-9-CM data files.
4. Mirth Connect routes the XDR message to the IHC NwHIN Aurion Gateway[10]
5. Aurion is an open source implementation of an NwHIN gateway. The gateway forwards the message using NwHIN protocols and services with secured two-way Transport Layer Security (TLS) with Security Assertion Markup Language (SAML) to the SHARPn NwHIN Aurion Gateway. This was implemented in a cloud service, residing behind the Mayo Clinic firewall.

6. The Aurion gateway on the SHARPn cloud passes the IHC XDR messages to Mirth Connect on the cloud.

6a. Mayo Clinic lab and ICD-9-CM data were generated as described for IHC above. These and clinical text documents are pushed to Mirth Connect on the SHARPn cloud behind Mayo Clinic's firewall.

7. Mirth Connect routes all incoming data to an Apache Unstructured Information Management Architecture (UIMA)[11] pipeline on the SHARPn cloud.
8. Normalization and NLP services are implemented in UIMA pipelines.
9. UIMA components route the normalized data back to Mirth Connect.
10. Mirth Connect routes the data instances to an open source SQL database, MySQL, where all data are persisted.
11. A rules engine queries the SQL database for normalized input data defined for a diabetes phenotyping algorithm.

The three main functions for the data normalization and phenotyping platform are described below:

2.2. EHR Data Normalization

2.2.1. Common Information Model

Clinical concepts must be normalized if decision support and analytic applications are to operate reliably on secondary EHR data. One EHR may store "resting heart rate" as a simple name-value pair (name = "resting heart rate" and value = "60 bpm"), while another EHR may store it in a more complex, *post-coordinated* fashion (name = "heart rate", value = "60 bpm", qualifier = "resting"). EHRs may contain more or less description or detail (*granularity*) about clinical events and observations. Further, the contents may be encoded using different terminologies, free text, or a mixture of both. A decision support application will need to take those differences into account when retrieving a clinical concept such as "resting heart rate", and *normalize* the data to a common representation before processing.

The Clinical Element Model (CEM) is Intermountain Healthcare's strategy for detailed clinical models.[12-19] A CEM provides a single logical structure for a data element, to which all incoming data of that type are normalized. CEMs are computable: they are written in the Constraint Definition Language (CDL) that can be compiled into objects that are passed between the layers of a system's technology stack. CEMs *do not dictate the physical implementation of those objects or of the stack components*. For example, CEMs may be compiled into XML Schema Definitions (XSDs), java or C# programming language classes, or Semantic Web Resource Description Framework (RDF) to provide structured model definitions in particular computing environments.[20] Correspondingly, CEM instances may be XML instances, java or C# objects, or RDF instances. CEMs, like HL7 version 2 and 3 models, prescribe that codes from controlled terminologies be used as the values of many of their attributes. Thus, CEMs facilitate interoperability between systems at a syntactic and semantic level while allowing for disparate architectures within those systems.

CEMs are agnostic of the origination of the data. Instances may be populated from textual reports by NLP services or from structured EHR data. All data stored in the EHR is encompassed in the scope of the CEM, including but not limited to:

- Allergies
- Problem lists
- Laboratory results
- Medication and diagnostic orders
- Medication administration
- Physical exam and clinical measurements
- Signs, symptoms, diagnoses
- Clinical documents
- Procedures
- Family history, medical history and review of symptoms.

Figure 2 shows a high-level abstraction of a medication order model and an instance of conformant data. The medication order CEM might prescribe that a medication order have attributes of “orderable item”, “dose”, “route”, and so on. Further, it may dictate that an orderable item’s value must be a code in the “orderable item value set” – a set list of controlled codes appropriate to represent orderable medications - and that the dose be represented as a “physical quantity” data type, which contains a numeric value and a code for a unit of measure, where that unit of measure code comes from a “dose unit” value set.

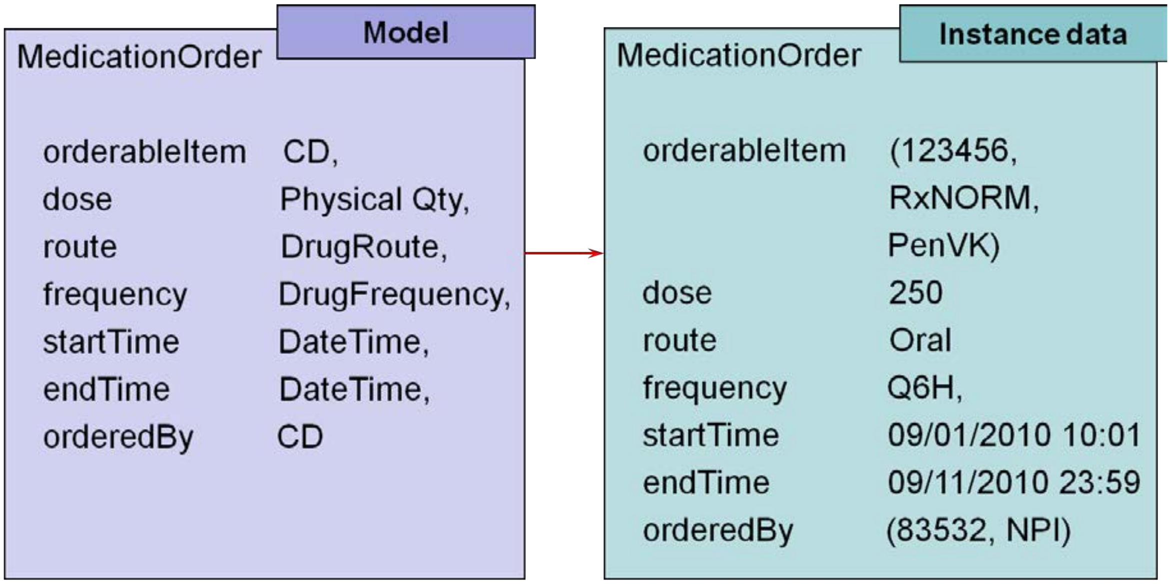

Figure 2. Partial CEM model and instance data for a medication order

2.2.2. Terminology Mapping Services

Structured, coded terminology plays a critical role in CEMs. In specifying the content of data instances, CEMs also specify the places in the instances where coded concepts must be used. The CEMs used by SHARPn will specify standard terminologies conformant with consolidated health informatics (CHI)[21]. Terminology services – especially mapping services – are essential to SHARPn success. An overview of the envisioned SHARPn normalization services are depicted in Figure 3.

SHARPn will align the terminology services it uses with the evolving Common Terminology Services (CTS) effort.[22] CTS defines a standard set of service interfaces for accessing and maintaining structured terminology. The second version of this effort (CTS2) is currently moving through the Object Management Group (OMG) standardization process.

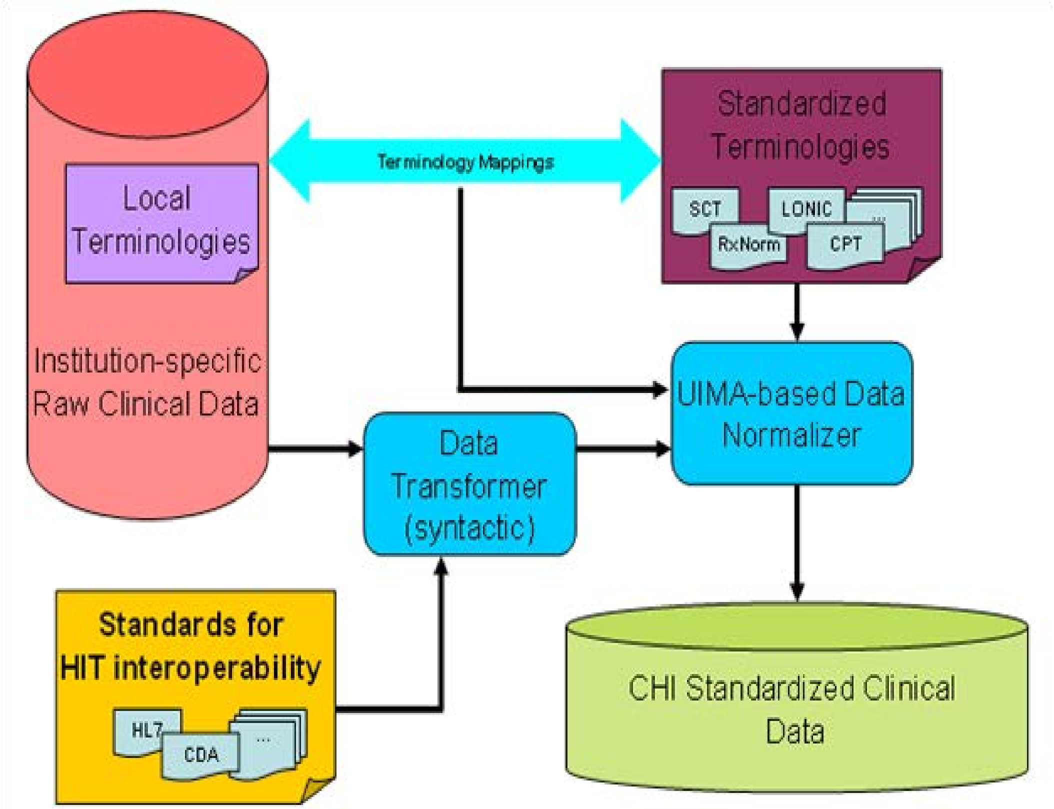
Local Terminologies
Institution-specific Raw Clinical Data
Terminology Mappings
Standardized Terminologies
SCT
LONIC
...
RxNorm
CPT
UIMA-based Data Normalizer
Data Transformer (syntactic)
Standards for HIT interoperability
HL7
CDA
...
CHI Standardized Clinical Data

Figure 3. SHARPn Normalization Services Overview

2.3. Natural Language Processing of Health Care Documentation Text

A substantial part of the information pertaining to a given patient is recorded in the unstructured, free-text part of the EHR. Information extraction (IE) is an application of

NLP aiming at converting the free-text into structured information. Normalization targets can vary from simple mappings to a terminology to more template-like structures. The CEMs provide such templates. Traditionally, IE and NLP have been applied to the domain of medicine for specific use cases where only a limited set of signals are to be discovered, such as discovering pneumonia from chest X-ray reports[23, 24] and peripheral arterial disease from radiology notes.[25] These approaches result in excellent performance but the application is limited to only the specific use cases for which they were developed. Our SHARPn philosophy takes a different approach. The goal is to discover a relevant set of summary information for a given patient in a disease- and use-case agnostic way, which is then stored and indexed for retrieval. The main principle, thus, is *process once, use multiple times*.

In the SHARPn project, the core NLP software is the clinical Text Analysis and Knowledge Extraction System (cTAKES).[26] As is true for the SHARPn platform, cTAKES is built within the Apache Unstructured Information Management Architecture (UIMA) which provides a solid basis for scalability, expandability and collaborative software development. One of the main UIMA concepts is that of a type system, which defines the annotations and their structure as generated by the pipeline. An agreed-upon basic type system ensures a software development foundation. The SHARPn NLP type system, which continues to be developed and evaluated, [27] implements the templates which are derived from the CEMs and represent the NLP normalization targets. The current cTAKES pipeline processes clinical notes and identifies the following clinical named entity mentions: drugs, diseases/disorders, signs/symptoms, anatomical sites, and procedures. The dictionary lookup algorithm uses a knowledge base as a resource created off SNOMED Clinical Terms® (SNOMED CT®)[28] and RxNorm[29] terms and their synonyms. Document input can be plain text or Clinical Document Architecture (CDA)[30] compliant XML documents.

2.4. High-throughput Phenotyping

The SHARPn view of phenotyping includes inclusion and exclusion criteria for clinical trials, numerator and denominator criteria for clinical quality metrics, epidemiologic criteria for outcomes research or observational studies, and trigger criteria for clinical decision support rules, among others. Our use implies the algorithmic recognition of any cohort within an EHR for a defined purpose. These purposes were inspired by the algorithmic identification of research phenotypes in the National Human Genome Research Institute (NHGRI) funded eMERGE (electronic MEdical Records and GEnomics) consortium[31] primarily to facilitate EHR-derived genomics studies. The underlying principles of using well-defined algorithms across diagnosis assertions, laboratory values, medication use, and NLP-derived observations adheres to the practices demonstrated in eMERGE[32].

Phenotyping algorithms in essence comprise a complex set of inclusion and exclusion criteria that are coupled using a set of logical operators. This lends to representing the algorithm criteria using a rules-based formalism. For SHARPn, we chose the open-source Drools production rule system.[33] Supported by the JBoss community, Drools comprises an inference engine, complex event processing and workflow system that provides a rich, declarative programming environment. The inference engine, also known as a “production rules system”, represents clinical knowledge as simple “conditions” and “actions”. The Drools system excels in that the rules engine and the workflow engine are tightly integrated and workflow can be used to visually represent complex relationships between logically related groups of rules. Development of a robust, common data model and scalable approaches to implement computable phenotyping algorithms within the Drools system are current SHARPn objectives.

The prototype SHARPn architecture includes persistence of the data. One can regard the persistence function as a data store for normalized representations. This is necessary because phenotyping constitutes a question/answer process, where within an algorithm one ascertains whether a particular patient does or does not manifest definitional criteria. This persistence layer can be realized by many data base technologies including: XML databases, object databases, RDF triple stores and SQL databases. For convenience in the development phases, we opted for SQL database technology. The ultimate goal is to decouple the algorithms from the physical implementation of the persistence layer.

3. METHODS: DATA THROUGHPUT DEMONSTRATION

After one year of design and development effort, a SHARPn platform was implemented for internal use. A demonstration of end to end throughput of high volume, real patient data was conducted in June, 2011. Approximately ten thousand patient records from two independent EHR systems, Mayo Clinic and Intermountain Healthcare were used to test the end to end flow of actual data. The use case was a common diabetes mellitus phenotyping algorithm. The algorithm was modified to normalized data specifications. The demonstration was focused on the acquisition and normalization of disparate data, and successful data flow from EHR to the population of input data in the rules engine.

The SHARPn platform infrastructure, software artifacts and research and development processes are designed and conducted with appropriate consideration of health information security and privacy requirements. Institutional review board policies are followed, and data are shared under appropriate data sharing agreements. In particular, site to site EHR data transmission for this demonstration involved exclusively de-identified data. Communication between the SHARPn organizations and the SHARPn cloud are secured in accordance with the NwHIN specifications. There is ongoing collaboration with the SHARP consortium tasked with EHR data security solutions.

3.1. Use Case – Diabetes Mellitus Phenotyping Algorithm

A phenotyping algorithm for a medicinally managed diabetes mellitus cohort was the use case for the data throughput demonstration. The algorithm was adapted from the eMERGE Northwestern Type 2 Diabetes Mellitus phenotyping algorithm[34] by the Southeast Minnesota Beacon Program[35] for purposes of group practice reporting.

A simplified algorithm (Figure 4) was derived to validate data throughput. The normalized EHR data to be used as input to the algorithm for this demonstration were:

- All ICD-9-CM codes in time period 1/1/09-12/31/10
- All ambulatory Medications ordered or noted as taken (RxNorm)
- Lab values (LOINC®[36]), results and date collected in time period 1/1/09-12/31/10

At least 2 face-to-face outpatient visits [1/1/09-12/31/10] with a diabetes ICD-9-CM code *[list of codes given]*

OR

Any medications ordered in [*list of brand and generic medication names]*

OR

At least 2 face-to-face outpatient visits [1/1/09-12/31/10] with a capillary glucose order in the measurement period OR with an abnormal blood glucose or HbA1c result [abnormal thresholds given], not glucose tolerance test and not part of pregnancy screen.

Figure 4. Simplified medicinally managed diabetes classification algorithm

3.2. Source EHR Data and Acquisition

3.2.1 Mayo Clinic

10,000 patients seen in 2010 at the Mayo Clinic were randomly selected from the Mayo research data warehouse, Enterprise Data Trust (EDT). The selection criteria were: (a) valid research authorization, (b) not deceased, (c) >=18 years as of June, 2011, (d) minimum of 2 visits in the last year with ICD-9-CM codes, (e) patient had at least 3 years of laboratory and medication data available. For selected patients, three years of prior clinical data were accessed, ending on June 30, 2010. The data sets selected were clinical notes, laboratory results and ICD-9-CM codes from the EDT. Unlike the de-identified Intermountain datasets, the Mayo patient data was processed locally and used to assess the platform's processing throughput and performance capabilities. No specific patient results were used or released.

Figure 5 depicts the data acquisition and flow. The cTAKES pipeline processed CDA formatted clinical documents from the Mayo Clinic. The focus for this demonstration was discovery of all noted medications relevant to the patient. The clinical notes of 10,000 patients from 2008-2010 were pulled from the Mayo EDT, resulting in 360,452 documents. Each document was then processed through cTAKES, matching RxNorm terms and their synonyms. This resulted in 3,442,000 Drug CEMs, which were passed to Mirth to deposit into the CEM database for phenotyping consumption.

The laboratory data was formatted into HL7 laboratory results messages using HL7 2.x syntax. The results were then normalized in a configurable UIMA pipeline. The ICD-9-CM data was extracted from the Mayo EDT and processed from tabular representations in text files by the AdminDiagnosis UIMA Processor.

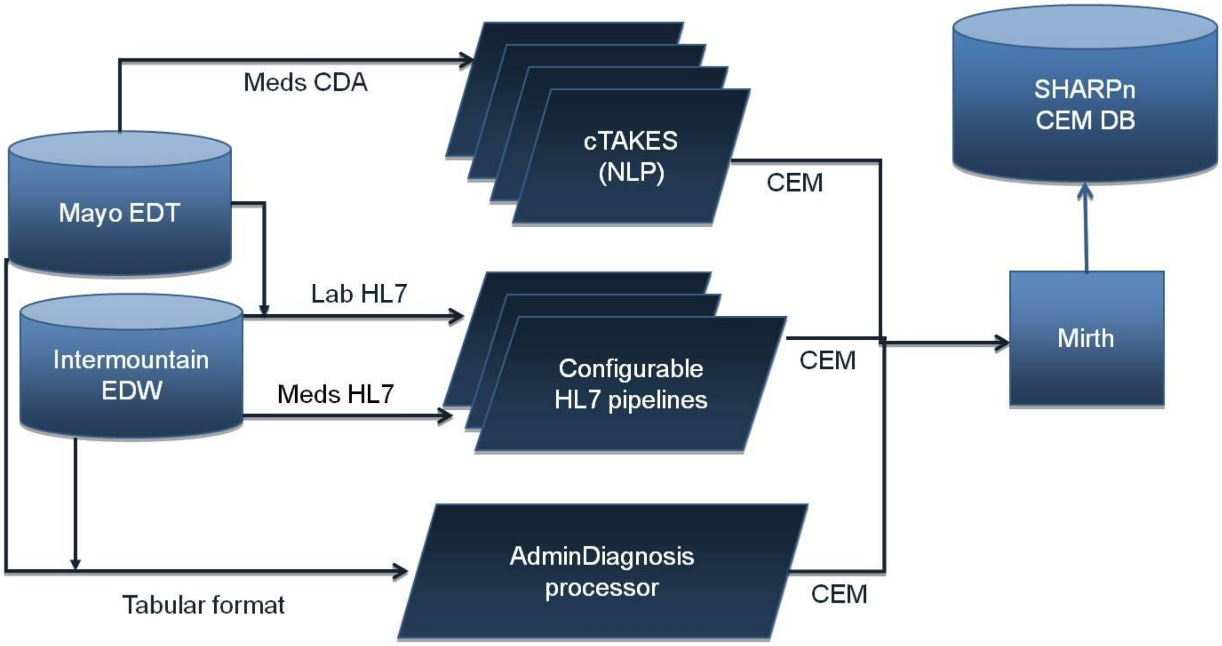

Figure 5. UIMA Normalization Processes accessed via Mirth

3.2.2 Intermountain Healthcare

One hundred cases were randomly selected from a curated pool of Intermountain test cases established for EHR development and validation activities. The test cases were developed from data stored in the Intermountain Enterprise Data Warehouse (EDW). The test data are real patient data, with all protected health identifiers of patients and providers removed and replaced with realistic false data. The encounter dates and other dates were shifted, using a random seed, so that the temporality of events is maintained. The test data were de-identified in compliance with HIPAA Guidelines.

All Medication List entries, ICD-9-CM codes with sequence numbers for the associated ambulatory encounters, and Intermountain clinical laboratory results for a period of time 7/1/2008 - 6/30/2011, to allow overlap with time period of interest for the classification algorithm, were extracted for the 100 test cases. Medication and laboratory data were prepared in HL7 2.7 format. This was accomplished by taking example Pharmacy

Order and Unsolicited Laboratory Result HL7 2.7 messages from the Intermountain production interfaces. Using these formats as templates, the false identifiers and true EDW data for the test cases were inserted into HL7 2.7 message and pushed to the framework where they were processed just as Mayo Clinic's. (Figure 5)

The Medication List may also contain text medication entries. For this demonstration, only coded medication entries were selected. The medications are encoded using First DataBank[37] (FDB) orderable drug codes (GCN). This code and the linked text description were sent in the HL7 input messages.

Although ICD-9-CM codes can be messaged in HL7 Diagnosis (DG1) segments, neither Mayo Clinic nor Intermountain Healthcare uses HL7 operationally for ambulatory ICD-9-CM codes. It was decided to prepare the ICD-9-CM data in a simple tab delimited, columnar text file. The unique ambulatory encounter identifiers, the shifted and de-identified encounter dates, the ICD-9-CM codes, their sequence numbers, and the provider service for the encounter were included. These data were processed like those from Mayo Clinic. (Figure 5)

3.3. Data Normalization Processes

3.3.1. Clinical Element Models

The data throughput demonstration required three CEMs:

- Ambulatory Medication Order
- Standard Lab Panel
- Administrative Diagnosis

These CEMs can be viewed at http://www.intermountainhealthcare.org/cem*.*

The CEMs were compiled into XSDs, which formed the basis of data instances processed by the SHARPn framework components. HL7 message structures, the tabular input files for administrative diagnosis codes, and the output of the NLP processes were mapped to these XSDs in UIMA Common Analysis Structure (CAS) components. UIMA CAS components render a common data structure for the application. Incoming HL7 and tabular data and NLP output were processed to populate XML instances that conformed to the CEM XSDs.

3.3.2. Terminology Mapping

The UIMA HL7 pipelines that processed structured medications and labs had components to translate site specific coded data to standardized terminologies. (Figure 6) Mayo internal lab codes and Sunquest codes from Intermountain were translated to LOINC codes. Intermountain FDB ordered medication (GCN) codes were translated to RxNorm Ingredient or Multiple Ingredient codes. We utilized a time tested and basic mechanism to achieve the translation step. The crosswalk tabular data (Mayo internal to LOINC, Sunquest Intermountain to LOINC and FDB GCN to RxNorm) were loaded into memory as lookup resources at initialization time, which then were used conditionally to perform the crosswalk at execution time, based on the origination of the data. These components served as proxies for the standards-based SHARPn terminology services currently in development.

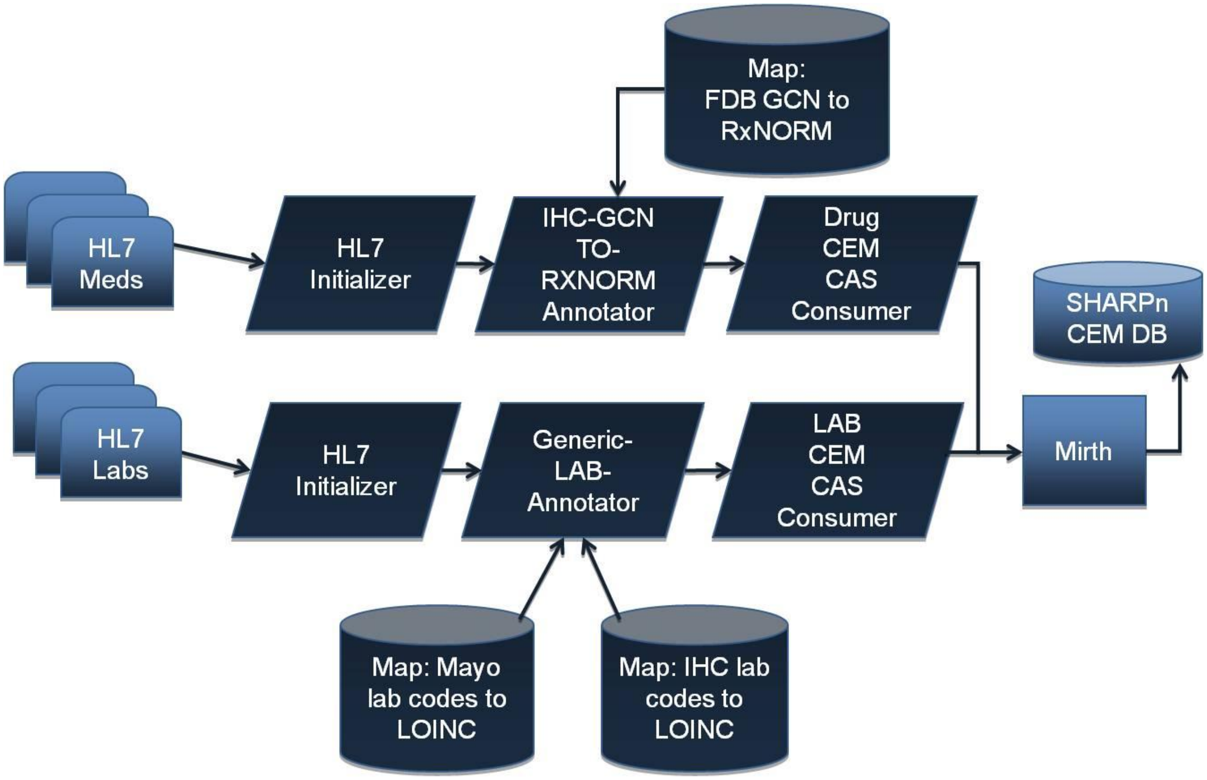

Figure 6. UIMA Terminology Mapping Processes

3.4 Data input to Diabetes Phenotyping Rules Implemented in Drools

Once the data were transformed and normalized to a consistent form, they were stored into a simple relational database, specifically open source MySQL. The CEM normalized data instances are represented as XML document instances. To facilitate access to that content, an index data record was added to support query access to stored clinical element data instances. Using these index records, the Drools data acquisition components were able to identify and access specific data stored in the normalized database.

The data throughput end goal was to populate a simple data model accommodating the three data types in the use case: ICD-9-CM codes, medications and laboratory results. The majority of data were successfully processed. Neither specific data transformation errors nor accuracy of the phenotyping rules were evaluated for this data throughput trial.

4 RESULTS AND DISCUSSION

The data throughput trial demonstrated successful functionality for three design goals: a secure, cloud-based transport architecture; a common syntax; and standardized semantics. These are requirements for interoperability of EHRs.[38] Both NLP-derived and structured data were normalized into common concept instances, which enabled a phenotyping algorithm to operate on patient data regardless of the concept's origin in the EHR. We believe the conceptual merging of structured and text-based EHR data is critically important to secondary use of EHR data as well as for patient data exchange for care purposes, decision support and the general case for retrieving a patient's information in the EHR. Recent studies have shown the value in using both text and structured data for more accurate disease classification.[39-41] The CEM's concept-based formalism unites them. The CEMs can be consumed at the concept level, or they can be constrained by their attributions if the researcher wishes to process NLP-derived versus structured source data separately.

SHARPn is focused on the generation of publicly available services and components to enable liquidity of standards-conforming, comparable, and consistent data for broad scale, high-throughput secondary use of the EHR. The architecture supports the definition and instantiation of terminology standards, as they are declared, into normalization components and services. The standardized data elements thus generated conform to a conceptual information model, and may be implemented in physical objects and physical information models as needed. We have described and demonstrated this functionality, and applied three terminology standards: ICD-9-CM, RxNorm, and LOINC. The implementation of this prototype platform and data throughput demonstration suggested challenges to our secondary EHR data liquidity goals:

4.1 Platform Architecture

A normalization and secondary use platform of the intended scalability requires a robust and secure architectural foundation. Data security, end to end connectivity and reliable data flow were critical to the normalization and phenotyping services. In the data throughput demonstration, architectural issues were discovered as expected. The data exchange and host processing platforms will require ongoing expertise from computer scientists with an understanding of health information exchange domain requirements.

4.2 Error Handling

Error checking and handling is a requirement for the framework. The prototype platform did not implement this functionality. Errors are expected to occur in multiple components. Some places that errors will occur are: (1) transmission related errors, (2) transformation errors, (3) EHR data quality errors, and (4) data persistence errors. Error recognition and proper handling will be highly important features of ubiquitous normalization and secondary use of EHR data.[38, 42]

4.3 Acquiring Source EHR Data

The pilot showed that HL7 2.x messages could be sent as the payload of an NwHIN Document Submission message, which could then be transformed and normalized using the UIMA pipeline. However, it is envisioned that there would be multiple formats for messages sent to the SHARPn cloud. HL7 messaging may cover the data that are typically interfaced between the EHR and ancillary systems, such as clinical lab, provider text documentation, and pharmacy. HL7 includes message structures that accommodate a variety of health information, such as diagnosis codes, allergies, and general clinical observations. But if organizations' operations do not include messaging those data, it is arguably no easier to acquire source data using a standardized HL7 representation than using other representations that can be mapped to CEMs.

The demonstration forced us to rethink the strategy of using HL7 Admit/ Discharge/ Transfer (ADT) messages to communicate ICD-9-CM codes. Since the organizations' business practices did not use HL7 to communicate ambulatory encounter codes, a text file format was populated by query at each organization and pushed to Mirth. Operationally, ICD-9-CM codes are messaged from providers to payors in ANSI standard X-12 messages[43], required under HIPAA for claims. These existing messaging processes and structure could be considered to harvest ICD-9-CM data in an automated manner. The acquisition of ICD-9-CM codes must consider their source. A practice management system may contain codes submitted or entered by providers, whereas other workflow processes may occur to select a subset of the codes or even modifications to submit as actual claims for billing purposes. These workflow processes differ among organizations, as can be stated about processes of populating EHR data in general.

There is untallied variation in the style, organization and formatting of clinical text documents among organizations and EHRs. The Mayo Clinic EHR environment implements a number of HL7 standards for clinical document organization. The CDA standard defines XML tags which mark the beginning and end of document sections. This unambiguous sectioning enables selective extraction of medications with variable levels of confidence depending on the section in which they occur in. A productive strategy is to limit the medication extraction to a set of sections such as Impression/Report/Plan, History of Present Illness, Current Medications, Admission Medications and Discharge Medications. The improvement in the accuracy of NLP information extraction methods when standardized clinical document styles are used remains a significant question. The current lack of standard styles in many clinical documents remains a challenge for normalization.

Design and development of SHARPn source EHR data acquisition tools require analysis of the origin, storage, and standard extraction or messaging mechanisms that may be tapped. In future iterations of the framework, Mirth Connect can take messages in many different forms, using many different protocols and transform them. SHARPn teams continue to develop acquisition strategies, focusing on the most useful EHR data categories. Meaningful use leaders and the National Quality Foundation have selected priority data types to support in EHRs and for quality measures.[44, 45]

4.4 Normalization Models

The CEMs used in the demonstration were based on the models created to populate EHR content. Based on feedback from the informatics community, SHARPn will instead design CEMs to cover general requirements for a spectrum of secondary use cases, realizing there will be needs for additional or different data requirements for particular uses. The clinical element models can be revised or extended as those needs arise. It is a challenge to create canonical models of EHR data for secondary use, as there are a variety of purposes to be served. However, large scale normalization of data for shared secondary use depends upon common models.

4.5 Terminology Services

The prototype platform did not implement runtime terminology services – simple table lookup mechanisms were implemented. Future SHARPn work will focus on calling standardized terminology services accessing a robust terminology server, as discussed in Section 2.2.2. Nevertheless, important insights regarding terminology mapping and services were gained during the pilot.

4.5.1 Mapping of Local Terminologies

A significant effort in the implementation of an EHR is mapping an institution's local codes to the terminologies stored and used within the EHR. Specialists or integration consultants are often required for the work. The SHARPn effort proved to be no exception as effort was required to complete the provider institutions' mappings from local laboratory observation codes to LOINC, even for the small demonstration project that was conducted. A large-scale data normalization project will need to take into account the considerable effort of mapping institutions' terminologies to standard terminologies. Compounding the problem is that many institutions will not be able to produce on demand the code sets sent in their messages in a form consumable by a terminology server, or at all. "Smart" tools are called for – tools that can analyze months-worth of an institution's messages and data and from them deduce the local code sets that need to be mapped to standards, and then even postulate the mappings.

Further, the inclusion of CEMs in the normalization solution suggests that not just terminology mapping solutions but model mapping solutions are required.

4.5.2 Medication Terminology

Mapping of medication codes is a particular type of mapping that introduces its own unique complexities. Healthcare provider institutions, vendors of drug information code sets, and the RxNorm standard all have multiple classes or hierarchies of codes, i.e., ingredients, generic drugs, branded drugs, and therapeutic classes. There are inevitably small differences in the structures of the parties' hierarchies and classes that cause mappings to be difficult. The mappings included with RxNorm realize much of the mapping work required between RxNorm and vendors; however, the work of mapping individual providers' codes and hierarchies to RxNorm will remain.

Furthermore, choices between hierarchies and classes need to be made. For example, Intermountain Healthcare chose to map its clinical drug data to one of the two different orderable drug code sets maintained by their drug information vendor. That class of drug code can map to RxNorm ingredient codes, RxNorm generic orderable drug codes, or RxNorm branded orderable drug codes. For the demonstration, we had to choose which of those to map to, and communicate to all parties involved: the team performing NLP, the team normalizing data, and the team executing phenotyping algorithms. The choices depend largely upon use cases. Decision support algorithms, for example, need to *know* when a patient is on any drug whose formulation contains a particular ingredient and are uninterested in the particular branding, formulations, dosage forms, or strengths of those drugs. Ordering applications, on the other hand, are very interested in brands, formulations, forms, and strengths. Given that different use cases demand different hierarchies of drug codes, it is inevitable and essential that support be provided to navigate between hierarchies.

4.5.3 Terminology Versioning and Updating

The SHARPn demonstration did not deal with multiple versions of terminologies or updates to terminologies, but it became apparent that any robust data normalization effort will need to do so. Controlled terminologies are not static, but undergo constant change as concepts are replaced by/redirected to other concepts, concepts are obsoleted, and new concepts are added. Generally, when an incoming code is mapped to a standard terminology, the latest changes in the standard terminology must be recognized and accessed through the mapping services. If a code has been made obsolete in the standard, a mapping to that code must be adjusted to point to the code that replaces the obsolete code and the services must navigate to the new code. If new codes are added to the standard, mappings must be updated and the services must access the new mappings in order to navigate to the newly added codes. If the meaning of a code changes, which occasionally occurs in some terminologies, mappings involving that code must be adjusted so as to continue to produce consistent results.

In some use cases, specific versions of terminologies must be recognized and accessed. Incoming codes may not always be from the latest version of a terminology, but they still need to be understood by the services and mapped correctly. Reporting requirements may dictate that certain versions of standard terminologies be sent out, requiring mappings from the internally-stored codes to various versions of outgoing terminologies. One example of this is found in standardization efforts such as Healthcare Information Technology Standards Panel (HITSP) and HL7 v3, which support the notion of dynamic and static bindings. In dynamic binding, a data element is bound to the current contents of a specified code set, e.g., the HITSP Procedure value set is bound to codes in the procedure hierarchy in whatever is the most current version of SNOMED CT. In contrast, in static binding, a data element is bound to a specific version of a code set, e.g., the HITSP Diagnosis Priority value set is bound to the 2.5.1 version of the HL7 Diagnosis Priority code system. Consequently, static binding requires terminology servers and services that maintain multiple versions of terminologies.

Terminology changes may occur frequently. RxNorm, for instance, is released monthly in an attempt to keep up with the frequently changing medication realm. Changing terminologies will clearly require considerable maintenance effort in order to make SHARPn data normalization and interoperability successful.

4.6 Specification of Standardized Phenotyping Algorithms

Another challenge highlighted by the demonstration was the difficulty of expressing this simple phenotyping algorithm's data specifications in terms of normalized objects and standard terminologies. Cohort identification algorithms' data are often expressed in abstract clinical terms, such as 'fasting glucose'. This expression maps to several LOINC codes. Existing data specifications must be translated to standardized data concept specifications. The computable specification of standardized phenotyping algorithm data and rules among them, are a recognized challenge with active engagement in the health informatics community.

5 CONCLUSIONS

A prototype platform was implemented from the developing SHARPn framework to perform secure transport, data normalization and common phenotyping services on disparate EHR data. A demonstration of data throughput from two large organizations was conducted to test the design and inform future development. The demonstration was successful in executing the intended end to end data processing flow, merging text and structured data into instances of clinical concepts, and normalizing disparate data to common objects with standard terminologies. We have shared the emerging architecture and observed challenges for standardization of EHR data for interoperable secondary use.

Acknowledgements

This manuscript was made possible by funding from the Strategic Health IT Advanced Research Projects (SHARP) Program administered by the Office of the National Coordinator for Health Information Technology. The contents of the manuscript are solely the responsibility of the authors.

The authors would like to thank the following individuals that assisted in this project or provided the computing infrastructure that made it possible:

Mayo Clinic
Jay Doughty
Sridhar Dwarkanath
Donna Ihrke
Vinod Kaggal
James Masanz

Childrens Hospital Boston
Pei Chen

Homer Warner Center for Informatics Research, Intermountain Healthcare
Jeffrey Ferraro
John Holmen, PhD
Herman Post
Darin Wilcox

We would also like to thank Lacey A. Hart, MBA, PMP®, Director, PMO Health Sciences Research, SEMN Beacon Program Manager, and SHARP Executive Officer, for providing administrative oversight and project management services.

References

[1] The Office of the National Coordinator for Health Information Technology, U. S. Department of Health and Human Services 2011 [cite August 10, 2011]; Available from: http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__home/1204.

[2] Strategic Health IT Advanced Research Projects (SHARP) Program, U. S. Department of Health and Human Services 2011 [cite Aug 28, 2011]; Available from: http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__sharp_program/1806.

[3] W.R. Hersh, Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance, Am J Manag Care 13 (2007) 277-278.

[4] C. Safran, M. Bloomrosen, W.E. Hammond, S. Labkoff, S. Markel-Fox, P.C. Tang, D.E. Detmer, P. Expert, Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper, J Am Med Inform Assoc 14 (2007) 1-9.

[5] C.G. Chute, J. Pathak, G.K. Savova, K.R. Bailey, M.I. Schor, L.A. Hart, C.E. Beebe, S.M. Huff, The SHARPn Project on Secondary Use of Electronic Medical Record Data: Progress, Plans, and Possibilities, American Medical Informatics Association 2011 Annual Symposium, Washington, D.C., 2011, pp. 248-256.

[6] V2 Messages, Health Level Seven International 2011 [cite Nov 1, 2011]; Available from: http://www.hl7.org/implement/standards/v2messages.cfm.

[7] ICD-9-CM Official Guidelines for Coding and Reporting The Centers for Medicare and Medicaid Services (CMS) and the National Center for Health Statistics (NCHS), 2009.

[8] Mirth Connect, Mirth Corporation 2011 [cite Nov 1, 2011]; Available from: http://www.mirthcorp.com/community/mirth-connect.

[9] Mapping the future of NwHIN, in: T. Sullivan, (Ed.), Government Health IT, MedTech Media, 2011.

[10] L. Westberg, Aurion Federal Gateway WSDL Interfaces, 2011 [cite Nov 1, 2011]; Available from: http://wiki.aurionproject.org/display/AurionMain/Aurion+Federal+Gateway+WSDL+Interfaces.

[11] Apache UIMA, The Apache Software Foundation 2011 [cite Nov 1, 2011]; Available from: http://uima.apache.org/.

[12] J.F. Coyle, A.R. Mori, S.M. Huff, Standards for detailed clinical models as the basis for medical data exchange and decision support, Int J Med Inform 69 (2003) 157-174.

[13] S.M. Huff, Practical modeling issues: Representing coded and structured patient data in EHR systems, AMIA 2010 Annual Symposium, Washington, D.C., 2010.

[14] S.M. Huff, R.A. Rocha, B.E. Bray, H.R. Warner, P.J. Haug, An event model of medical information representation, J Am Med Inform Assoc 2 (1995) 116-134.

[15] S.M. Huff, R.A. Rocha, J.F. Coyle, S.P. Narus, Integrating detailed clinical models into application development tools, Stud Health Technol Inform 107 (2004) 1058-1062.

[16] S.M. Huff, R.A. Rocha, H.R. Solbrig, M.W. Barnes, S.P. Schrank, M. Smith, Linking a medical vocabulary to a clinical data model using Abstract Syntax Notation 1, Methods Inf Med 37 (1998) 440-452.

[17] T. Oniki, A Detailed Clinical Model Development Environment, 12th International HL7 Interoperability Conference, Orlando, Florida, US, 2011.

[18] C.G. Parker, R.A. Rocha, J.R. Campbell, S.W. Tu, S.M. Huff, Detailed clinical models for sharable, executable guidelines, Stud Health Technol Inform 107 (2004) 145-148.

[19] D.J. Steiner, J.F. Coyle, B.H. Rocha, P. Haug, S.M. Huff, Medical data abstractionism: fitting an EMR to radically evolving medical information systems, Stud Health Technol Inform 107 (2004) 550-554.

[20] C. Tao, C.G. Parker, T.A. Oniki, J. Pathak, S.M. Huff, C.G. Chute, An OWL Meta-Ontology for Representing the Clinical Element Model, American Medical Informatics Association Annual Symposium (2011).

[21] Office of the National Coordinator for Health Information Technology (ONC) Presidential Initiatives: Consolidated Health Informatics U. S. Department of Health & Human Services 2006 [cite Aug 15, 2011]; Available from: http://www.hhs.gov/healthit/chiinitiative.html.

[22] Common Terminology Service 2, Mayo Clinic 2011 [cite Aug 15, 2011]; Available from: http://informatics.mayo.edu/cts2.

[23] L.M. Christensen, P.J. Haug, M. Fiszman, MPLUS: a probabilistic medical language understanding system, Proceedings of the ACL-02 workshop on Natural language processing in the biomedical

domain - Volume 3, Association for Computational Linguistics, Phildadelphia, Pennsylvania, 2002, pp. 29-36.

[24] M. Fiszman, W.W. Chapman, D. Aronsky, R.S. Evans, P.J. Haug, Automatic detection of acute bacterial pneumonia from chest X-ray reports, J Am Med Inform Assoc 7 (2000) 593-604.

[25] G.K. Savova, J. Fan, Z. Ye, S.P. Murphy, J. Zheng, C.G. Chute, I.J. Kullo, Discovering peripheral arterial disease cases from radiology notes using natural language processing, AMIA Annu Symp Proc 2010 (2010) 722-726.

[26] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J Am Med Inform Assoc 17 (2010) 507-513.

[27] S. Wu, V. Kaggal, G. Savova, H. Liu, D. Dligach, J. Zheng, W. Chapman, C. Chute, Generality and reuse in a common type system for clinical natural language processing, Managing Interoperability and Complexity in Health Systems (MIXHS 2011) in conjuction with the 20th ACM International conference on information and knowledge management, Glasgow, UK, 2011.

[28] Unified Medical Language System® (UMLS®) SNOMED Clinical Terms® (SNOMED CT®), U.S. National Library of Medicine 2011 [cite Jan 17, 2012]; Available from: http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html.

[29] Unified Medical Language System® (UMLS®) RxNorm, U.S. National Library of Medicine 2011 [cite Jan 17, 2012]; Available from: http://www.nlm.nih.gov/research/umls/rxnorm/.

[30] R.H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F.M. Behlen, P.V. Biron, A. Shabo Shvo, HL7 Clinical Document Architecture, Release 2, J Am Med Inform Assoc, 2006, pp. 30-39.

[31] C.A. McCarty, R.L. Chisholm, C.G. Chute, I.J. Kullo, G.P. Jarvik, E.B. Larson, R. Li, D.R. Masys, M.D. Ritchie, D.M. Roden, J.P. Struewing, W.A. Wolf, The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies, BMC Med Genomics 4 (2011) 13.

[32] I.J. Kullo, J. Fan, J. Pathak, G.K. Savova, Z. Ali, C.G. Chute, Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease, J Am Med Inform Assoc 17 (2010) 568-574.

[33] Drools Business Logic integration Platform, JBoss Community 2011 [cite Aug 15, 2011]; Available from: http://www.jboss.org/drools.

[34] eMERGE, eMERGE Network Phenotype Library, 2011 [cite 13 March 2011]; Available from: https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Library_of_Phenotype_Algorithms.

[35] Southest Minnesota Beacon Program, 2011 [cite Aug 10, 2011]; Available from: http://semnbeacon.wordpress.com/.

[36] Logical Observation Identifiers Names and Codes (LOINC®), Regenstrief Institute, Inc. 2012 [cite Jan 17, 2012]; Available from: www.loinc.org.

[37] First DataBank, [cite Jan 17, 2012]; Available from: www.firstdatabank.com.

[38] W.E. Hammond, C. Bailey, P. Boucher, M. Spohr, P. Whitaker, Connecting information to improve health, Health Aff (Millwood) 29 (2010) 284-288.

[39] L. Li, H.S. Chase, C.O. Patel, C. Friedman, C. Weng, Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study, AMIA Annu Symp Proc (2008) 404-408.

[40] K.P. Liao, T. Cai, V. Gainer, S. Goryachev, Q. Zeng-treitler, S. Raychaudhuri, P. Szolovits, S. Churchill, S. Murphy, I. Kohane, E.W. Karlson, R.M. Plenge, Electronic medical records for discovery research in rheumatoid arthritis, Arthritis Care Res (Hoboken) 62 (2010) 1120-1127.

[41] H.J. Murff, F. FitzHenry, M.E. Matheny, N. Gentry, K.L. Kotter, K. Crimin, R.S. Dittus, A.K. Rosen, P.L. Elkin, S.H. Brown, T. Speroff, Automated identification of postoperative complications within an electronic medical record using natural language processing, JAMA 306 (2011) 848-855.

[42] K.S. Chan, J.B. Fowles, J.P. Weiner, Review: electronic health records and the reliability and validity of quality measures: a review of the literature, Med Care Res Rev 67 (2010) 503-527.

[43] ASC X12 The Accredited Standards Committee, 2011 [cite Aug 15, 2011]; Available from: http://www.x12.org/.

[44] Quality Data Model, National Quality Forum 2011 [cite Aug 15, 2011]; Available from: http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx.

[45] Standards & Certification, Office of the National Coordinator for Health Information Technology 2011 [cite Aug 15, 2011]; Available from: http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__standards_and_certification/1153.