Office of the National Coordinator for Health Information Technology

Strategic Health IT Advanced Research Projects (SHARP)

AREA 4: Secondary Use of EHR Data (SHARPn) Program

Annual Progress Report

Reporting period: 1/1/2012 – 12/31/2012

Program: AREA 4 - Secondary Use of EHR Data (SHARPn)
Award Number: 90TR0002
Prime DUNS: 006471700
Principal Investigators: Christopher Chute, MD, DrPh, Mayo Clinic;
Stan Huff, MD, Intermountain Healthcare
Program Manager: Lacey Hart, MBA, PMP®

n n n

The AREA 4 - Secondary Use of EHR Data (SHARPn) is one of four Strategic Healthcare IT Advanced Research Projects (SHARP) grantees funded by the U.S. Department of Health and Human Services (HHS), through the Office of the National Coordinator for Health Information Technology.

The SHARP Program is a federally funded project to address well-documented problems that impede the adoption of health IT. The knowledge generated and innovations created from this program will accelerate progress toward the meaningful use of health IT and a high-performing, adaptive, nationwide health care system.

The purpose of this annual report is to provide ONC and other Government officials with information about the progress made against planned grantee activities.

Demonstrated in 2012, SHARPn has developed and deployed a compilation of open-source and publicly available applications and resources that support universal exchange, sharing and reuse of EHR data collected during routine clinical processes as part of patient care.

In this Annual Report you will find updates to development and research advances in three main themes: (1) framework for normalization and standardization of clinical data—both structured and unstructured data extracted from clinical narratives, and includes clinical information modeling and terminology standards (2) platform for representing and executing patient cohort identification and phenotyping logic, and (3) evaluation of data quality and consistency.

Table of Contents

Program Background

AREA 4 - Secondary Use of EHR Data (SHARPn) is a collaboration of 14 academic and industry partners to develop tools and resources that influence and extend secondary uses of clinical data. The project will enhance patient safety and improve patient medical outcomes through the use of an electronic health record. Traditionally, a patient's medical information, such as medical history, exam data, hospital visits and physician notes, are stored inconsistently and in multiple locations, both electronically and non-electronically.

Area four's mission is to enable the use of EHR data for secondary purposes, such as clinical research and public health. By creating tangible, scalable, and open-source tools, services and software for large-scale health record data sharing; this project will ultimately help improve the quality and efficiency of patient care through the use of an electronic health record.

The program proposed to assemble modular services and agents from existing open-source software to improve the utilization of EHR data for a spectrum of use-cases and focus on three themes: Normalization, Phenotypes, and Data Quality/Evaluation. The program was operationalized into six projects that span one or more of these themes, though together constitute a coherent ensemble of related research and development. Finally, these services will have open-source deployments as well as commercially supported implementations.

The six projects are strongly intertwined, mutually dependent projects, including: 1) Semantic and Syntactic Normalization 2) Natural Language Processing (NLP) 3) Phenotype Applications 4) Performance Optimization 5) Data Quality Metrics and 6) Evaluation Frameworks. The first two projects align with our Data Normalization theme, with Phenotype Applications and Performance Optimization spanning themes 1 and 2 (Normalization and Phenotyping); while the last two projects correspond to our third theme (Data Quality/Evaluation).

Collaborators:

- Agilex Technologies
- Clinical Data Interchange Standards Consortium (CDISC)
- Centerphase Solutions
- Deloitte
- Seattle Group Health
- IBM Watson Research Labs
- University of Utah
- Harvard University/Childrens Hospital Boston
- Intermountain Healthcare (IHC)
- Mayo Clinic
- MIT
- MITRE
- State University of New York, Albany
- University of Pittsburgh
- University of Colorado
- University of California, San Diego

SHARP Area 4 Announcements can be found at the following URL: www.sharpn.org

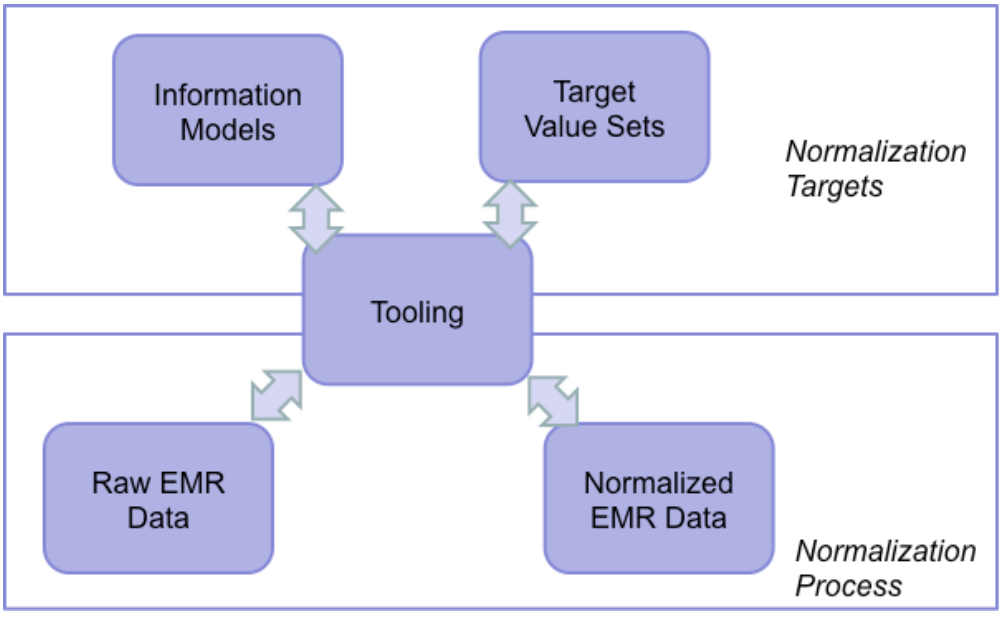
Clinical Data Normalization (DN)

Data normalization and clinical models are at the heart of secondary use of clinical data. If the data is not comparable and consistent between sources, it can't be aggregated into large data sets and used for example to reliably answer research questions or survey populations from multiple health organizations.

Detailed clinical models are the basis for retaining computable meaning when data is exchanged between heterogeneous computer systems. Detailed clinical models are also the basis for shared computable meaning when clinical data is referenced in decision support logic.

- The need for the clinical models is dictated by what we want to accomplish as providers of health care
- The best clinical care requires the use of computerized clinical decision support and automated data analysis
- Clinical decision support and automated data analysis can only function against standard, structured, and coded data
- The detailed clinical models provide the standard structure and terminology needed for clinical decision support and automated data analysis

DN Aims: To conduct the science for realizing semantic interoperability and integration of diverse data sources; to develop tools and resources enabling the generation of normalized Electronic Medical Record (EMR) data for secondary uses.



DN Progress Overview:
The Data Norm team has progressed in their work with clinical model needs from a variety of perspectives as well as releasing modularized normalization tools.

The work within the Data Normalization project this year revolved around four foci:

1) Data Models and Value Sets
   a. identifying use cases and performing gap analysis of models and value sets with identified use cases
2) Standards
   a. alignment of models and value sets with national/international standards adopted in Meaningful Use and/or ONC S&I framework
3) Usability (Tools)
   a. modularize different pieces involved in data normalization pipeline (i.e., modeling, value sets, pipeline implementation, and persistent store) to de-couple the dependency among them
4) Accessibility
   a. releasing tools and resources to the community early and frequently

Data Models and Value Sets:
For the first focus, as in previous years, the SHARPn team has leveraged the Clinical Element Model (CEM) originating from Intermountain Health Care (IHC) to represent detailed clinical data models. CEMs are the basis for retaining computable meaning when data is exchanged between heterogeneous computer systems. They are also the origin for shared computable meaning when clinical data is referenced in decision support logic.

Work this year continued to evaluate these models against several clinical and research use cases. Early in the year there was significant progress in creating different models for different use cases including 'Core' models for: patient demographics, medications, laboratory data, administrative diagnosis and procedure, and disease/disorder. Definition of value sets in terms of standards was leveraged where possible. Applications allow one to leverage attributes from and across different models to focus on specific topics/characteristics. There is also ongoing effort to develop, improve, and achieve common terminologies, mapping models, and natural language processing applications to support data collection, data extraction, and manipulation for specific research/problem targets.

Lessons learned mid-year were captured for the SHARPn Annual meeting in June 2012: http://informatics.mayo.edu/sharp/index.php/Annual_Gathering

Highlights of these 'lessons learned' include

1) "One model fits all" is difficult to achieve with multiple models out there:
   a. Clinical Trials (e.g., CDISC CSHARE) vs Secondary Use (e.g., SHARPn)
   b. Proprietary EMR (e.g., GE Qualibria) and Open Secondary Use (e.g., SHARPn)
2) The root of all modeling questions: precoordination vs. postcoordination and what to store in the model instance vs. leave in the terminology

The Clinical Element Models developed for SHARPn are GE Healthcare copyrighted and licensed to Intermountain Healthcare to distribute freely under an open content license agreement found at www.clinicalelement.com.

Through continued collaboration with the SHARPn High Throughput Phenotyping (HTP) team as well as Mayo Clinic's Pharmacogenomics Research Network (PGRN) team, we performed additional gap analysis of these Core/Secondary Use models, and identified the need for additional CEM models which capture Allergy, Vital Signs, Signs and Symptoms, and Problems. The need for representing semantic relationships among different CEM instances was also revealed.

Specifically, the Mayo Clinic PGRN/SHARPn working group conducted standardization efforts to collect data dictionaries for 4483 variables across PGRN sites. All the studies in PGRN have a variety of data representations. The working group took meta-data and ran through data-preprocessing into a SHARPn centralized database. It mapped components and semantic annotations and categorized variables into: demographics, disease disorder, laboratory, medication, clinical observations, and smoking status. The result was that categorized variables could indeed be mapped to SHARPn CEMs. Results from the study indicated that 54% of the Pharmacogenomics variables were able to be mapped to CEMs, demonstrating that this method can be used to normalize genomics study data dictionaries.

A key component in defining and instantiating the CEMs are structured codes and terminologies. In particular, the CEMs used by SHARPn adopt Meaningful Use standards and value sets for specifying diagnosis, medications, lab results, and other classes of data. Terminology services – especially mapping services – are essential, and we have adopted Common Terminology Services 2 (CTS2) as our core infrastructure. More in-depth information regarding our CTS2 efforts can be found later in this report.

Standards

For the second focus, the SHARPn team continues to support and engage in an international context to addressing clinical models with the Clinical Information Modeling Initiative (CIMI), whose vision is improve the interoperability of healthcare systems through shared implementable clinical information models. This large initiative benefits the outcomes of the work being done on SHARPn, and therefore, the team has discussed the ability of the infrastructure to be CIMI-ready. The full CIMI Reference Model consists of:

- Core Reference Model
- Data Type Value Types
- Demographics Model
- Supporting Classes

To maximize the reusability of the CEMs in a variety of use cases across both clinical study and secondary use, it was necessary to build interoperability between the CEMs and existing data standards (e.g. CDISC and ISO 11179 standards). A sub-project formed a CSHARE CEMs Harmonization Working Group with representatives from CDISC, Intermountain Healthcare and Mayo Clinic with the objective to harmonize SHARPn

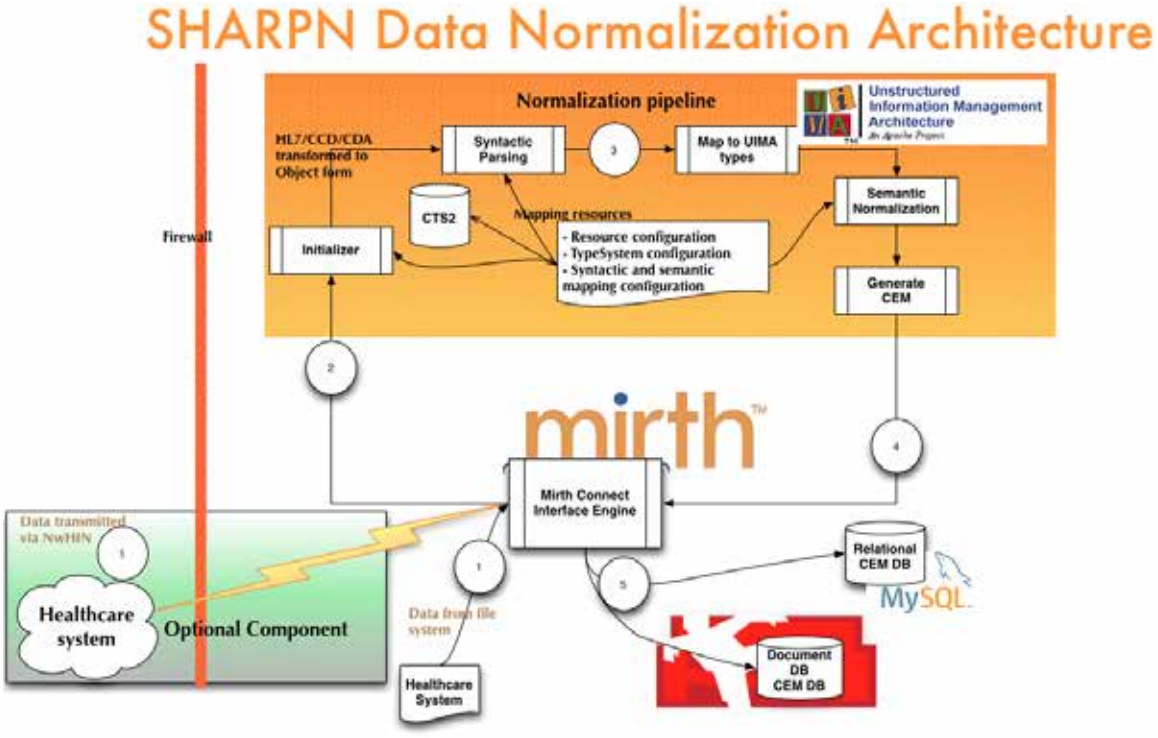
Clinical Element Models with Clinical Data Interchange Standards Consortium (CDISC) Shared Health and Clinical Research Electronic Library (SHARE) Clinical Study Data Standards.

As the starting point, the team focused on three generic domains: Demographics, Lab Tests and Medications. The working group performed a panel review on each data element extracted from the CDISC templates and SHARPn CEMs. When a consensus was achieved, a data element was classified into one of the following three context categories: Common, Clinical Study or Secondary Use. In total, they reviewed 127 data elements from the CDISC SHARE templates and 1130 data elements extracted from the SHARPn CEMs. In conclusion, we have identified a set of data elements that are common to the context of both clinical study and secondary use. We consider that the outcomes produced by this working group would be useful for facilitating the semantic interoperability between systems for both clinical study and secondary use.

The SHARPn team has also been participating in the S&I framework and have led HL7 document architecture so that both models and value sets proposed by SHARPn will be aligned with standards adopted by the S&I framework and HL7 standards.

Usability (Tools)

The third focus area, usability, concentrates on modularizing components involved in data normalization. Modular components will increase the level of usability by being a part of a greater, more flexible infrastructure. The components can be represented as four sections: modeling, value sets, pipeline implementation, and persistence store. The team has developed and implemented a new data normalization architecture which supports this third focus.



The SHARPn clinical data normalization pipeline (diagram, above) consists of five main sections/processes: (1) Data to be normalized is read from the file system. This data can also be transmitted on NwHIN via TCP/IP from an external entity. (2) Mirth Connect invokes the normalization pipeline using one of its predefined channels and passes the data/document (ex: HL7, CCD, tabular-data) to be normalized. (3) Normalization pipeline goes through initialization of the components (including loading resources from the file system or other predefined resource such as CTS2 service) and then performs syntactic parsing and semantic normalization to generate normalized data in the form of CEM. (4) Normalized data is handed back to Mirth Connect. (5) Mirth Connect uses one of the predefined channels to serialize the normalized CEM data to CouchDB or MySQL based on the configuration.

More information regarding the development of the normalization pipeline can be found under the Infrastructure portion of this report.

Accessibility

The fourth focus area, accessibility, is important to the user community, enabling access to the most recent releases available to them, post-development. In order to release artifacts early and often, we have developed a web site (http://www.clinicalelement.com) to release CEM models for various projects (including SHARPn CEM models) in diverse formats. The web site supports the ability to search and browse different CEM models in a graphical tree view. It also supports change request functionality where a user may create new requests, view requests made by his/her team, and modify/add information to requests he/she has made. This tool facilitates the communication process and allows the Intermountain Healthcare modeling team to capture all requests in a single database.

Notable Milestones Reached in DN:

1) Launched Clinical Element Model (CEM) browser/request site
2) Re-engineered SHARPn data normalization pipeline to be resource-driven so that source models, target models, value sets, syntactic and semantic mappings can be achieved through external configuration
3) Adopted a generic XML parser developed previously to make the model mapping generic
4) Released models, value sets, and pipeline for the following objects: demographics (SecondaryUsePatient), medication (SecondaryUseNotedDrug), lab observations (SecondaryUseLabObservation), diagnosis (AdministrativeDiagnosis), procedures (AdministrativeProcedure), disease/disorder (SecondaryUseAssertion)
5) Processed HL7 messages, Continuity of Care Document (CCD) documents, and Clinical Document Architecture (CDA) documents to generate SecondaryUsePatient and SecondaryUseNotedDrug CEM instances for Pan-SHARP activities
6) Initial round of validation of SecondaryUsePatient and SecondaryUseNotedDrug generated from Intermountain Healthcare data is complete. We are in the process of addressing the questions that came out of the first round of evaluations.
7) Terminology value sets define the valid values used in the models
8) Fostered CTS2 participation in CIMI effort
9) SHARPn Value Sets published as CTS2 Resolved Value Sets

a) ICD-9
b) RxNorm
c) LOINC
d) ECIS
e) SNOMED-CT

10) CTS2 Services Created For
    a) SNOMED-CT Simple Refsets and Dynamic Value Set Definition
    b) ECIS Value Sets and Maps
11) Updated each of the CEM to SQL Mirth Channels to support the latest 2012 CEM structures for use by Data Normalization and HTP teams
12) Completed the investigation of adopting CouchDB, an open source document database system, for storing CEM instances
13) Configured Mirth Channels for the implementation of CouchDB-based CEM DB
14) Developed prototype APIs to access CouchDB-based CEM DB populated with CEM instances of 400 patients for phenotyping and evaluation teams.

DN Next Steps:

1) Continue development of the CEM website (requesting and browsing) to support display of mindmaps, value sets, and model sub-parts and hierarchical tree navigation
2) Release of additional, defined models, including clinical-based procedure(SecondaryUseProcedure)
3) Pipeline development
    a) Integration of NLP modules into normalization pipeline
    b) GUI to facilitate easier creation and maintenance pipeline configuration
    c) Configure the pipeline to generate Signs/Symptoms CEMs
    d) Incorporate CTS2 standards with value sets and embed into the CEM Browser
    e) Value sets can be modified by removing entities or by adding entities from external CTS2 compliant repositories.
4) CTS2 Services to be created for
    a) UMLS
    b) BioPortal
    c) CIMI
5) The CEM to CouchDB work will continue into 2013 with the addition of several new CEM Structures (Allergies, Procedures, etc.) along with some ongoing work with the Validation Functions that were developed under Release 1. This will culminate with Release 2 of the CEM to CouchDB. The work related to developing scripts to administer CouchDB database will continue.
6) Solidify data type values from CEMs (XML, DB, RDF, OWL, Triple Store, SWRL, CCDA).
7) Mapping analysis of models and value sets between Consolidated Clinical Document Architecture (CCDA) meaningful use requirements and SHARPn in order to interface with the SMART team

Clinical Natural Language Processing (cNLP)

The clinical narrative within the EHR consists primarily of physician and nurse notes describing the patient's status, disease or tissue/image. The knowledge contained in unstructured textual documents (e.g., pathology reports, clinical notes), is critical to achieving all of these goals. Its normalization requires methods that deal with the variety and complexity of human language.

cNLP systems process the free text clinical narrative to discover information from textual reports and make it searchable. One clinical application is to search for a given diagnosis and perhaps to find common co-morbidities associated with it; all these are relevant clinical events which when summarized give the full and detailed profile of patients' histories.

cNLP Aims:
Information extraction (IE): transformation of unstructured text into structured representations and merging clinical data extracted from free text with structured data

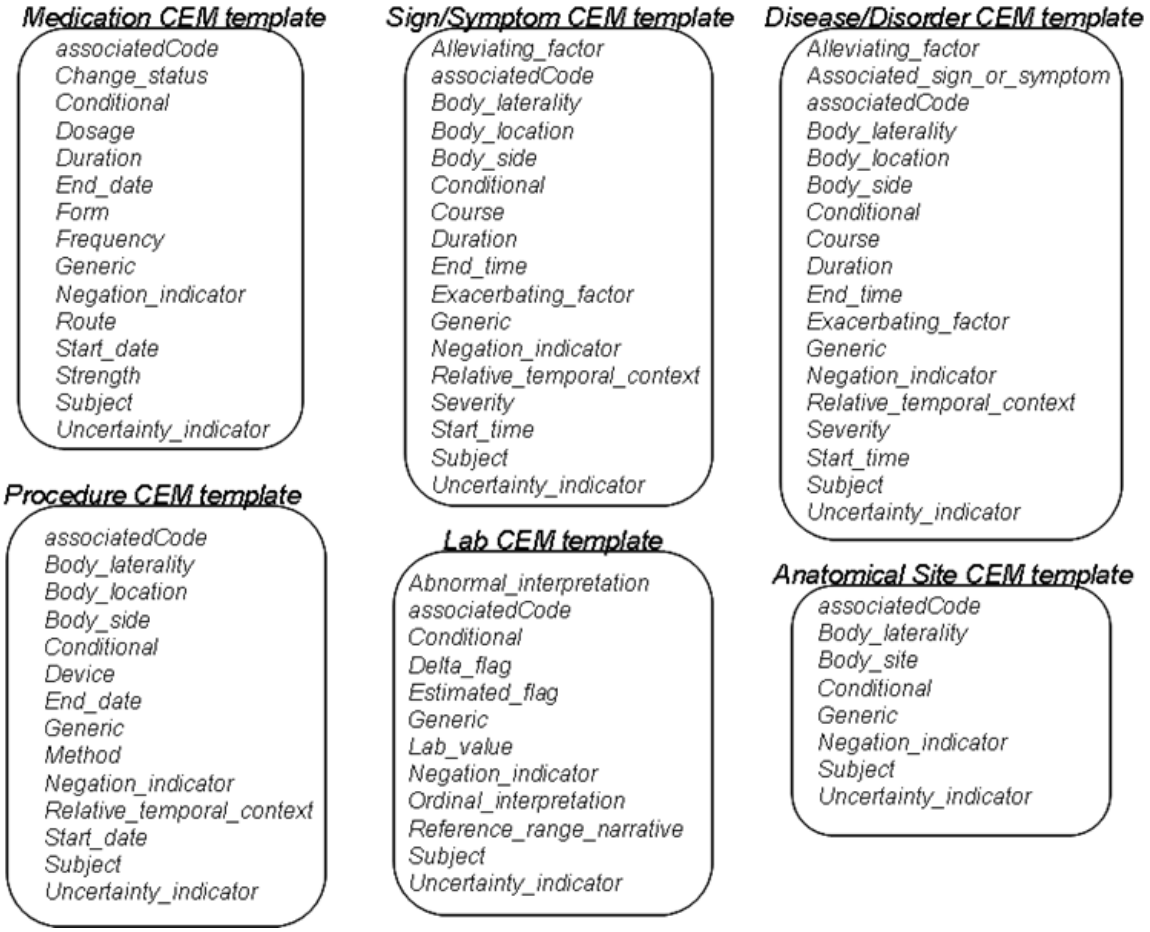
- Entity and Event discovery
- Relation discovery
- Normalization template: Clinical Element Model (CEM)

Overarching goal

- high-throughput phenotype extraction from clinical free text based on standards and the principles of interoperability
- general purpose clinical NLP tool with applications to the majority of all imaginable use cases

Progress:
The clinical narrative within the EHR consists primarily of physician and nurse notes describing the patient's status, disease or tissue/image. Its normalization requires methods that deal with the variety and complexity of human language. We have developed sophisticated IE methods to discover a relevant set of normalized summary information for a given patient in a disease- and use-case agnostic way.

We have defined six "templates" – abstractions of CEM noted in the data normalization section – which are populated by processing the textual information and then are mapped to the models. The anchors for each template are a Medication, a Sign/Symptom, a Disease/Disorder, a Procedure, a Lab and an Anatomical Site respectively.
Some attributes are relevant to all templates, for example negation_indicator, others are specific to a particular template, for example dosage is specific to Medications.



The main methods used for the normalization of the clinical narrative are rules and supervised machine learning (ML). As with any supervised ML techniques, the algorithms require labeled data points to learn the patterns from as well as to evaluate the output. Therefore, a corpus representative of the EHR from two SHARPn institutions (Mayo Clinic and Seattle Group Health) was sampled and de-identified following HIPAA guidelines. The de-identification process was a combination of automatic output from MITRE Identification Scrubber Tool (MIST; http://mist-deid.sourceforge.net/) and manual review. The corpus size is 500K words that we intend to share with the research community under Data Use Agreements with the originating institutions. The corpus was annotated by experts for several layers following carefully extended or newly developed annotation guidelines conformant with established standards and conventions in the NLP field to allow inter-operability. The annotated layers (syntactic and semantic) allow learning structures over increasingly complex language representations thus enabling state-of-the-art IE from the clinical narrative. A detailed description of the syntactic and select semantic layers is provided in Albright et al. 2013.

The best performing methods are implemented as part of the Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) which is built using UIMA. It is comprised of a variety of modules including sentence boundary detection, syntactic parsers, named entity

recognition, negation/uncertainty discovery, and coreference resolver, to name a few. Additional details on cTAKES are available via http://incubator.apache.org/ctakes/.

The cTAKES software is now part of the Apache Software Foundation (http://incubator.apache.org/ctakes/). The move from SourceForge to the well-known Apache Software Foundation paves the road for international adoption as well as contributions. In addition cTAKES now incorporates ClearTK (http://code.google.com/p/cleartk/), the machine learning NLP from University of Colorado, one of the SHARP cNLP contributing sites.

Two main sets of software activities define the contributions of the SHARPn NLP team. The first set of activities is the development of new cTAKES functionalities aiming at extracting and creating a synthesis of the full picture of a patient's clinical events. The second set of activities focuses on engineering the software to facilitate large-scale robust processing and enthusiastic adoption by the community be it research or industry. Thus, the cTAKES product is envisioned to apply to 90% of all use cases in biomedicine – clinical research, phenotyping, meaningful use, comparative effectiveness, patient care, decision support systems, phenotype/genotype association studies, etc.

Towards the goal of developing new cTAKES functionalities to accommodate the broad range of use cases, the SHARPn cNLP team accomplished the following:

1) The creation of gold standard annotated corpus for use for developing and evaluating machine learning components. This is an impactful contribution especially given the fact that gold standard follows standards and conventions adopted by the broader NLP community (Penn Treebank, Propbank, UMLS). Annotated gold standard data for the period – Treebank (90K words), Propbank (207K words), Coreference (53K words), UMLS entities and their respective CEM templates (86K words), UMLS relation (53K words).
2) The development of a new web-based tool, Anafora, for better support of the gold standard development of clinical notes. Release of the tool open source.
3) New, improved models were developed for part-of-speech tagging, dependency parsing, and semantic role labeling as part of the open source ClearNLP project. The OpenNLP/ClearParser part-of-speech tagger, dependency parser and semantic role labeler, trained on Mayo data, were released in May 2012 as part of cTAKES 2.5. Evaluation is described in Albright et al, 2013 (see list of publications)
4) Implementation of named entity recognition system (MedTagger) and evaluation through the participation of the 2012 i2b2 NLP challenge
5) The curation of MedLex (a semantic lexicon for clinical NLP)
6) Developed, implemented and tested several unsupervised algorithms for entity disambiguation, using the UMLS concept mappings. Implemented and tested approaches include (1) unsupervised knowledge based methods, including path-based distance algorithm using SNOMED hierarchical relations between concepts and personalized PageRank algorithm using concept relations from the UMLS Metathesaurus, and (2) unsupervised bottom-up sense induction algorithms using bag-of-words, syntactic dependency features, including feature pattern induction methods that adapt two topic modeling algorithms, Latent Dirichlet Allocation and Hierarchical

Dirichlet Process, to the problem of sense induction. Algorithm performance was tested on the Mayo word-sense disambiguation corpus

7) Version 1 of negation, uncertainty, and conditional attribute annotator. Conversion of MITRE's assertion status module to an annotator for negation, uncertainty, and conditional attributes. There were 6 assertion (*present, absent, possible, hypothetical, conditional, not patient*) were mutually exclusive. Version 1 of the negation, uncertainty, and conditional attribute annotator retained the statistical model trained on i2b2 data and converted the output of that model to attributes in SHARPn's type system representation. It associated i2b2 *absent* with **polarity** = *-1*, i2b2 *possible* with **uncertainty** = *1*, and i2b2 *hypothetical* with **conditional** = *true*. Its functionality was extended to assign attribute values not only to disease/disorders but also to medication events, anatomical locations, and procedures.
8) Refactoring the assertion engine with ClearTK. Benefits of this refactoring include (1) making it easier to build features by using high level API rather than low level UIMA, (2) making it simpler to train the model with different feature sets, training parameters, and data sets, and (3) making the module easier to adapt to different NLP pipelines
9) Implementation of a series of ClearTK classifiers (both training and inference) for attributes, including subject and generic. These now include task-specific rule-based logic in their features. Evaluation of the modules.
10) Development of methods and their evaluation for the discovery of bodySite and severity attributes. Evaluation was done on the SHARP gold standard corpus. Performance is comparable to that of human annotators.
11) Developed new medication extraction/normalization tool (MedER)
12) Updated Apache cTAKES Drug NER to align with the new common type system
13) Improved Apache cTAKES Drug NER capability to correctly link drug with attributes and change status determination
14) Improvement of the methods for coreference resolution - incorporating wikipedia features into the system by reducing the size of the index from ~27GB to ~36MB as well the output of the MIT sectionizer. New joint features which represent complex but important coreference phenomena. This resulted in a substantial improvement in performance. Additional features were developed that are combinations of other features that are surprisingly powerful.
15) Strategy for creating instance-based templates
16) Beta of template population delivered to infrastructure team. This was built on cTAKES 2.5
17) Migration of cTAKES code to apache.org
18) Completed setting up build process for cTAKES at apache.org
19) NLP evaluation workbench - added the ability to view statistics on attribute/value assignments associated with document annotations, in the same way that users can view statistics associated with classification assignments.
20) New machine learning sectionizer - The work completed over the last year has focused on analyzing clinical narrative reports to obtain a better understanding of the structural variance across document types and institutions. Further we have begun the annotation of the seven most frequent report types: consultation reports, discharge

summaries, operative reports, surgical pathology reports, history & physical reports, emergency room reports, and x-ray reports.

Notable Milestones Reached in NLP:

1) Gold standard annotations – Treebank (257K words), Propbank (206K words), Coreference (53K words), UMLS entities and CEM templates (86K words), UMLS relations (53K words)
2) Annotated clinical narratives for structural decomposition to assist in building the new machine learning sectionizer
3) The MedTagger has been implemented using the Apache cTAKES common type system and will be available in public after manuscript acceptance.
4) Testing of different frameworks for entity disambiguation is complete, with unsupervised bottom-up methods producing best results
5) Version 1 of negation, uncertainty, and conditional attribute annotator completed and released
6) Version 2 of negation, uncertainty, and conditional attribute annotator (ClearTK-based implementation) developed and in testing phase
7) First-pass, rule-based implementation of subject and generic modules completed and released
8) Second-pass, ClearTK-based implementation of subject and generic modules in testing phase
9) The fully functional relation module is available for download as part of Apache cTAKES. The released module includes the models trained on the SHARP data. This includes the models for the discovery of the bodySite and severity modifiers
10) Apache cTAKES Drug NER refinement
11) New medication extraction/normalization tool (MedER)
12) Refined coreference resolution module with performance improvement on a dataset of Mayo Clinic notes by 10-15 percentage points
13) Beta of template population delivered to infrastructure team. This was built on cTAKES 2.5
14) NLP Evaluation workbench can take Knowtator annotations as input, and can compare input from different sources, e.g. UIMA pipeline output compared with gold standard annotations in Knowtator format.
15) Incorporation of the MIT sectionizer into Apache cTAKES
16) New Machine learning sectionizer - completed annotation guidelines for structural decomposition and developed annotation schema. Completed initial annotation of 500 reports of each report type

NLP Next Steps:

1) Integrating the improved ClearNLP models into cTAKES
2) Re-training using SHARPn data sets
3) Evaluation of assertion annotator accuracy using available annotated Seed Corpus data. Retraining with additional annotated data as it becomes available
4) Comparing assertion status annotator's performance to the performance of the cTAKES negation detection annotator, which is a pattern-based approach (no MaxEnt

models required/used) that uses finite state machines and is roughly based on the popular NegEX. The comparison would be done using the seed corpus for evaluation
5) Training models for subject and generic attribute discovery on SHARP annotations to evaluate performance. Error analysis and feature engineering based on corpus analysis.
6) Exploring joint inference for negation, uncertainty, conditional, subject, and generic
7) The addition of the code to evaluate the relation extraction module within an end-to-end system framework
8) Error analysis to come up new ways of improving system performance for the relation module including the discovery of bodySite and severity attributes of clinical events
9) Full evaluation and refinements of MedER
10) Alignment of MedER with the Apache cTAKES common type system
11) MedER system release as a module within Apache cTAKES
12) Implementation of new features for coreference resolution to capture semantic similarity, using wikipedia/medpedia knowledge, upstream components such as assertion and SRL, and knowledge resources. Incorporation in the Apache cTAKES 3.1 release (Q1 2013)
13) In discussions with the Apache OpenNLP project about combining efforts in a single coreference module shared between the projects and trained independently for general domain or medical domain.
14) The UIMA type system is being updated to allow further work on the CEM extraction system (template populations)
15) Migrate beta code for instance-based template population into latest version of Apache cTAKES
16) Refinements of the instance-based template population module
17) Apache cTAKES (incubating) 3.10 release – planned for 1Q2013
18) Improve Apache cTAKES end user install experience
19) More automated tests for Apache cTAKES
20) The NLP Evaluation Workbench will be embedded in a UIMA analysis engine, allowing it to be run at the endpoint of a UIMA pipeline without generating intermediate result files. We are planning to allow users to create complex queries involving logical combinations of classification and attribute/value assignments, on which to match annotations. We will add the ability to input annotation sets in CSV format, and explore other formats, as well as the possibility of a Java API which a user can extend for viewing annotations with arbitrary formats and characteristics
21) New machine learning sectionizer - Section header identification approach using concept extraction aggregated with machine learning classification. Section header identification using a cluster matching. Section header identification using Latent Dirichlet allocation with topic change identification

High-throughput Phenotyping (HTP)

Phenotyping is identifying a set of characteristics of about a patient, such as: A diagnosis, Demographics or A set of lab results. A well-defined phenotype will produce a group of patients who might be eligible for a clinical study or a program to support high-risk patients.

While originally for application of research cohorts from EMR's, this project has obvious extensions to clinical trial eligibility, clinical decision support and has relevance to quality metrics (numerator and denominator constitute phenotypes). HTP leverages high-throughput computational technologies to derive efficient use of health information data. The field currently has barriers in technological research and tool generation.

HTP Aims:
To develop techniques and algorithms that operate on normalized EMR data to identify cohorts of potentially eligible subjects on the basis of disease, symptoms, or related findings. The project is developing:

- Phenotyping processes
- Algorithms for specific diseases
- Tools to incorporate data from multiple sites

Progress:

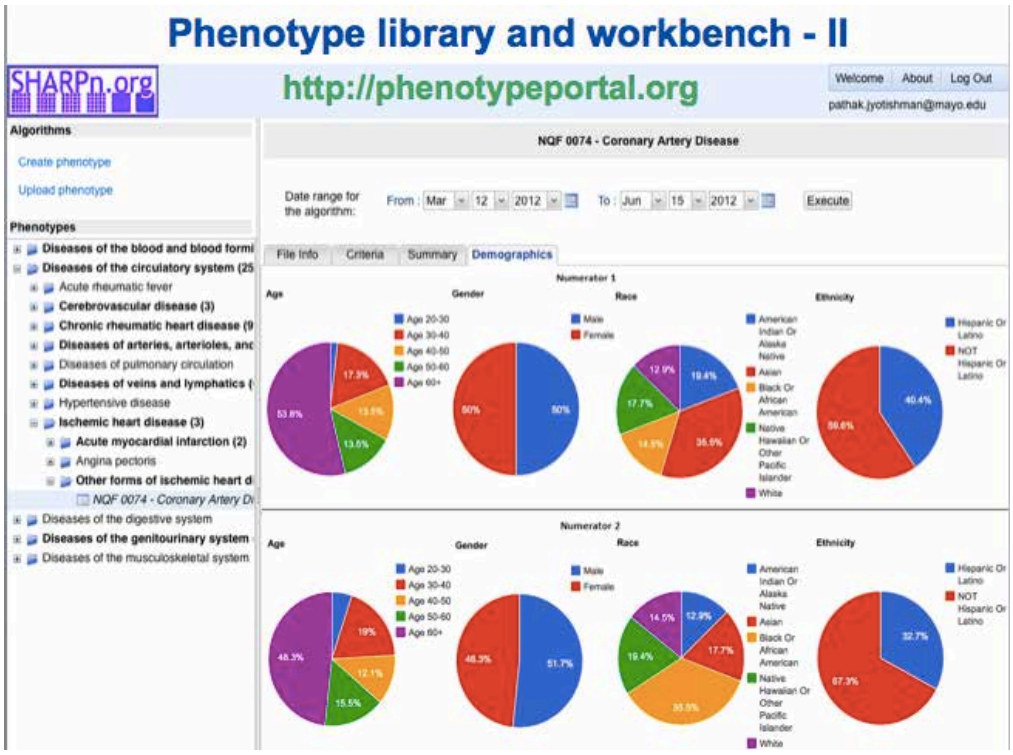
HTP Web Application
In early 2012, PopHealth (http://projectpophealth.org) and Smart GWT (http://www.smartclient.com/) were both considered and evaluated as a visual interface for algorithm execution. After careful review, Smart GWT was selected and a prototype was quickly implemented to demonstrate how easily a rich user interface could be created to interact with phenotype data. Smart GWT allowed the team to write code in Java, and via the Google Web Toolkit, the Java code was translated into JavaScript so that it could run in common browsers without the need for any plugins. This provided the same look and feel on all major clients running a browser.

Upon acceptance of the prototype, the team developed additional features and enhancements. Special consideration was given to using Standards within the tooling. Most recently, the team has implemented the value set functionality of the OMG Common Terminology Services 2 (CTS2) Standard. This allowed the team to represent the algorithm criteria value sets in a CTS2 service that persists the value set content.

Major features of the phenotype portal:

- Portal registration and email notifications.
- Ability to upload NQF files (algorithms) to the phenotype portal and persist them in a database.

- View the algorithm criteria. The value set criteria is being served from a CTS2 value set service.
- Ability to execute algorithms in real time and return results.
- Communication to the execution server is done via Representational State Transfer (REST) calls.
- Execution results are displayed with an integrated charting tool package, HighCharts.



The portal is publicly available on the Internet at http://www.phenotypeportal.org/. Architecturally, all Internet requests that go to http://www.phenotypeportal.org/ are redirected to Mayo Clinic's public cloud infrastructure where the HTP server application is deployed.

The client and server code for the phenotype portal is open source and available on SVN at the following public URL: https://svn.code.sf.net/p/sharpn/htp/code

API Implementation for querying CEM DB and Investigation of SMART APIs
In order to allow for a more flexible build and deployment procedure, new build and packaging scripts were added to the Quality Data Model (QDM) to Drools Converter code. This allows for distribution of a single executable artifact, which simplifies deployment. This also allows for easier integration into automated test environments, which will ensure testability going forward.

A new interface layer was also added to decouple the QDM to Drools Converter logic from the data extraction logic. This has enabled us to consider multiple data sources as inputs to the Converter -- one of which was the CEM DB. Interfaces were successfully implemented

to allow the QDM to Drools Converter to extract data from the CEM DB.

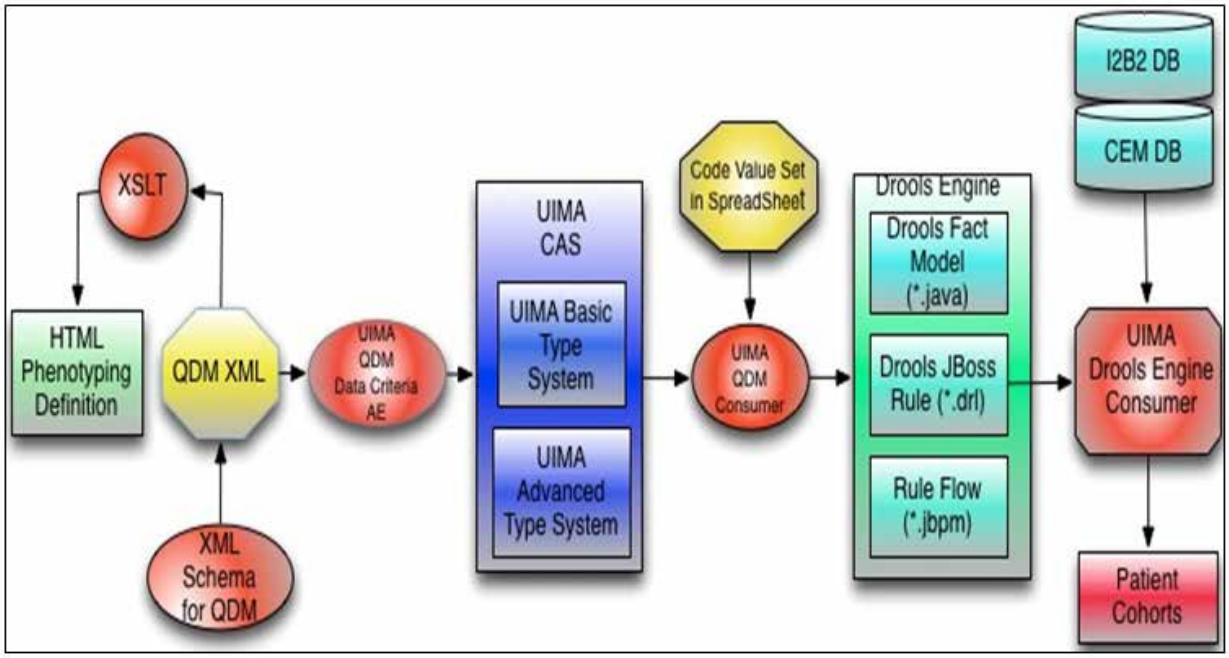
With this interface layer in place, we are positioned to explore other datasources, one of which being SMART containers. Investigation of the SMART API is ongoing — we have installed a reference SMART container and created a client to exercise the SMART API. With these artifacts in place, we can begin analyzing the API in terms of gaps in functionality between it and what the CEM DB data can provide.

Semantic Web representation of CEMs and querying
We have represented the CEM using semantic web notations. The CEM is an information model designed for representing clinical information in EHR systems across organizations. The current representation of CEMs does not support formal semantic definitions and therefore it is not possible to perform reasoning and consistency checking on derived models. Our project focuses on representing the CEM specification using the Web Ontology Language (OWL). The CEM-OWL representation connects the CEM content with the Semantic Web environment, which provides authoring, reasoning, and querying tools. This work may also facilitate the harmonization of the CEMs with domain knowledge represented in terminology models as well as other clinical information models such as the openEHR Archetype Model. We have created the CEM-OWL meta ontology based on the CEM specification. A convertor has been implemented in Java to automatically translate detailed CEMs from XML to OWL. A panel evaluation has been conducted, and the results show that the OWL modeling can faithfully represent the CEM specification and represent patient data.

Once we have the CEM-OWL model ready, patient data can be represented using Resource Description Framework (RDF) through the SHARPn EHR data normalization pipeline. To apply phenotyping algorithms or quality measures to the patient data, however, requires a machine executable, semantic compatible representation of the algorithms and measures. To achieve this objective, we have prototyped the representation of National Quality Forum (NQF) quality measures using the Semantic Web Rule Language (SWRL) to allow automatic execution of the performance measure criteria to normalized EHR data in RDF. An Application Programming Interface (API) has been implemented to compose SWRL rules based on the NQF measure criteria specification. Using the API, we have transformed three NQF measures to SWRL rules. The evaluation result on a simulated EHR triple store indicated that our API can successfully compose SWRL rules for the NQF measures and retrieve eligible patients based on the criteria.

Drools-based execution of phenotyping algorithms
To electronically execute phenotype algorithms, SHARPn has set up a JBoss open-source Java-based application server and Drools Business Logic integration platform which provides a unified and integrated platform to run the phenotype Rules, Workflow and Event Processing in the SHARPn cloud. Preliminary representation of the medicinally managed diabetes algorithm and hypothyroidism prototype was completed using Drools. More specifically, we have constructed an UIMA pipeline to reach this goal. The pipeline--called Qdm2DroolsTranslator—enables the following:

1. Parsing NQF algorithms, which are represented as xml files and converting NQF data criteria and population criteria into UIMA CAS, the common data type system and the data structure of UIMA.
2. Further converting CAS-represented data criteria and population criteria into JBoss Drools and Java fact models. In the conversion, we use CEM models as the mapping goals. Namely, mapping from NQF-standard to CEM-standard is implemented as well in this process. Drools scripts are composed of individual medical events as patient.drl, lab.drl, medication.drl, diagnosis.drl, encounter.drl, physicalExam.drl and and so on. Each drool script is supposed to check whether patients are eligible for meeting some data criteria defined by NQF. The fact models are in the form of java objects.
3. Extracting patients from Mayo CouchDB with their diagnosis information and then launching drools scripts against them to discover patient diseases' status.
4. Identifying patient cohorts by building population criteria trees and finally outputting the cohorts with patient demographics.



The translator has been tested with multiple QDM eMeasures on simulated data and one eMeasure (NQF-64) on true diabetes patient data with satisfactory results. In future work, we will update the translator following the updates from NQF. Meanwhile, we will test our translator on more eMeasures and with more patient data.

Mapping between NQF QDM and CEMs

The SHARPn collaboration is developing services to translate and execute Meaningful Use (MU) eMeasures, expressed in the QDM. QDM includes a model for specification of data elements and a grammar for the logic or rules that comprise an eMeasure (http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx). QDM data elements were mapped to normalized EHR data, conformant to SHARPn CEMs. CEMs

enable the mapping and extraction from disparate EHRs to a normalized representation of the source data. The SHARPn CEMs are bound to standard value sets specified for MU. The 'key' code, an identifier for a specific clinical observation and the 'data' code, an identifier for the coded value of an observation, are generally the CEM 'qualifiers' that enable a mapping from a QDM data element specification to one or more CEMs. QDM data specifications may include attributes of provenance, such as date/time, and other category specific attributes, such as the order status of a medication. These are also accommodated in CEMs. Our goal was a computable mapping between the QDM data elements and CEMs for the subsequent development, validation and deployment of eMeasures in the SHARPn platform, which is intended to support multiple EHRs.

The first step we took toward a computable mapping was to evaluate the coverage of existing SHARPn CEMs and identify gaps to resolve with the help of the SHARPn Normalization Team. Approximately 1,262 data elements across 45 MU Stage 1 eMeasures were considered for initial conceptual mapping. In process, the Stage 2 eMeasures and a summary list of data elements were published: 12 eMeasures and approximately 213 data elements that were not included in Stage 2 were eliminated. Data elements that were not compatible with EHR data content according to MU Stage 2 EHR certification requirements were also excluded in our initial mapping. The remaining data elements were collapsed into specifications of QDM 'category' plus 'state' plus 'coding system(s)'. For example: the specification, category 'DIAGNOSIS', state 'ACTIVE', and coding systems 'ICD-10-CM, ICD-9-CM, SNOMED-CT', covers many disease specific QDM data elements across eMeasures. Another example is category 'DIAGNOSTIC STUDY', state 'RESULT', and coding system 'CPT'. In a series of reviews between the Phenotyping and Normalization teams, these derived QDM specifications were mapped to existing or additional CEMs planned and subsequently implemented as a result of this conceptual phase of mapping. We are refining these conceptual mappings according to the newly released MU Stage 2 eMeasure data specifications.

The next step toward mapping the QDM data elements to CEMs was to generate detailed mappings from QDM data elements in specific Stage 2 eMeasures to use in development and validation of the SHARPn eMeasure execution services. In this phase, we have also generated detailed mappings from the CEM data required for the eMeasures to the HL7 V2.5 information models currently used in SHARPn to transport source data to the normalization pipeline. These detailed mappings were available or generated with help from the Evaluation and Infrastructure Teams. We are continuing to map, validate and document mappings for specific eMeasures, with the goal of a computable QDM to CEM mapping that enables the execution of all of the MU Stage 2 eMeasures on multiple and disparate sources of EHR data.

Infrastructure & Scalability

Building modules, resources, pipelines, and their local or network instantiations does not always scale in high-demand, high-throughput circumstances. Ensuring that these services can divide clinical streams into coherent tasks, access shared standards repositories, and leverage parallel and cloud computing infrastructures requires a separate level of software engineering.

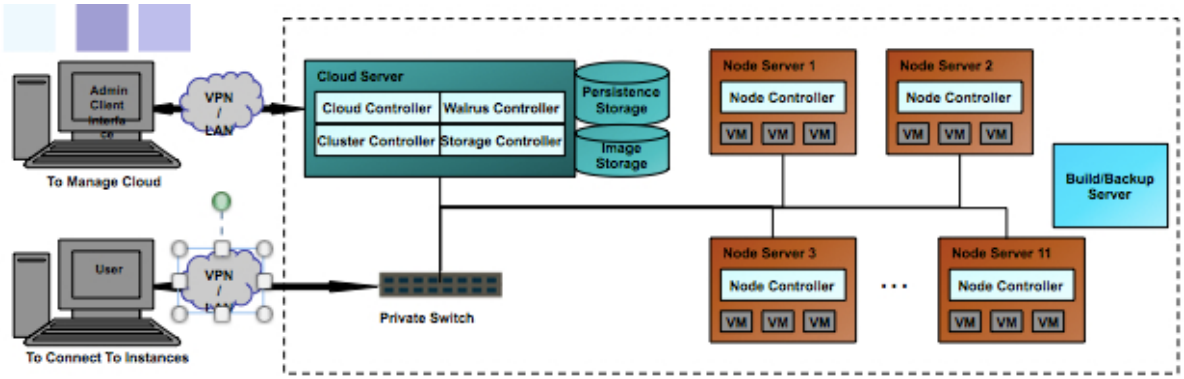
Infrastructure Aims:

1) Parallelize services to run on clusters
2) Allow interprocess communication among multiple "clones" of servers
3) Develop and deploy virtual machine images that can dynamically scale in cloud computing environments.

Progress:

Cloud Computing Environment

The SHARPn cloud computing environment, also known as "the cloud" has been maintained by the SHARPn infrastructure team and is basically a set of virtual machine images that has been instantiated, used, and shut down for Secondary Use of EHR Data research. The file share server has about 7TB of space currently set up.



With the release of the cloud, SHARPn has generated end-user documentation on the base cloud infrastructure and content is maintained on the Cloud wiki site: http://informatics.mayo.edu/cirruswiki/index.php/Project_SHARPN

Normalization Pipeline

The year of 2012 found the Infrastructure team collaborating more closely than ever with the other SHARPn teams, as they diligently strive to ensure development follows all necessary requirements. Infrastructure worked in tandem with Data Normalization, which has yielded a significant and positive impact on the ability to deliver.

Building modules, resources, pipelines, and their local or network instantiations does not always scale in high-demand, high-throughput circumstances. Ensuring that these services can divide clinical streams into coherent tasks, access shared standards repositories, and leverage parallel and cloud computing infrastructures requires a separate level of software engineering. This project ensures that these software resources can scale to near-real-time performance or operate as standalone applications.

The SHARPn data normalization pipeline adopts Mirth, an open source health care integration engine, as the interface engine and the Apache Unstructured Information Management Architecture (UIMA) as the platform. It takes advantage of Mirth's ability to support the creation of interfaces between disparate systems and UIMA's resource configuration ability to enable the transformation of heterogeneous EHR data sources (including clinical narratives) to common clinical models and standard value sets.

The behavior of the normalization pipeline is driven by several variables besides input and output, all handled by UIMA's resource reconfiguration:

1. Source information models
2. Target information models
3. Source value sets
4. Target value sets
5. The mapping between source and target models (syntactic normalization)
6. The mapping between source and target value sets (semantic normalization).

As discussed in the Data Normalization section, we adopt the CEMs and CTS2 infrastructure for defining the information models and access to terminologies and value sets, respectively, for enabling syntactic and semantic normalizations.

In particular, syntactic normalization specifies where (in the source data or other location) to obtain the values that will fill the target model. They are "structural" mappings - mapping the input structure (e.g., an HL7 message) to the output structure (a CEM).

There are five mapping types defined:

- Constant (C): The target CEM field is a constant. (In this case, the Source Root Element and Source xPath are ignored.) For example, SourceDataFormat is a constant.
- One to one mapping (S): one value from the source instantiates a value only in one CEM instance
- One to many mapping (M): one value from the source instantiates a value in multiple CEM instances
- Inference (I): the value that needs to be populated in a CEM field is inferred from another field (the Source Root Element and Source xPath will be acquired from the target CEMs rather than the incoming source data)
- Conditional Inference (X): inferences are based on conditional logic

For semantic normalization, we draw value sets for problems, diagnoses, lab observations, medications and other classes of data from Meaningful Use terminologies hosted in a local

CTS2 server instance. The entire normalization process relies on the creation or identification of mapping files that map to these terminologies.

Data Transfer/Processing: We have primarily leveraged two industry standard open source tools for the data exchange and transfer: Mirth Connect and Aurion (NwHIN). Aurion provides Gateway to Gateway data exchange using the NwHIN XDR (Document Submission) protocol enabling participating partners to push clinical documents in a variety of forms (HL7 2.x, CDA, and CEM). Mirth Connect software and channels have been developed to provide Sender, Receiver, Transformation, and Persistence channels in support of the variety of use cases as well as enabling the interconnectivity of the various SHARPn systems.

Data Storage/Retrieval: The SHARPn teams have developed two different technology solutions for the storage and retrieval of Secondary Use data. The first is an open source SQL-based solution (channels, software, and SQL data models) for the processing and storage of each of the SHARPn Secondary Use data models. This solution enables each individual CEM instance record to be stored in a standardized SQL data model. Data can be queried directly from the SQL Secondary Use SHARPn tables and extracted as individual fields or as complete XML records. The second is an open source document storage solution that leverages the open source "CouchDB" repository to store the CEM XML data as individual JSON documents providing a more document centric view into the SHARPn data. JSON documents can be extracted and converted back to XML documents as required. Both solutions leverage toolsets built around the CEM data models for a variety of clinical data including NotedDrugs, Labs, Administrative Diagnostic, and other clinically relevant data.

Over the course of 2012, the end-to-end pipeline normalization Pipeline was used for several tests to refine the infrastructure and to support the Pan-SHARP project discussed in a later section.

To demonstrate the applicability of our tools and infrastructure, we experimented with phenotyping algorithms from eMERGE and PGRN consortiums, as well as MU eMeasures. Currently the team has successfully run 10 algorithms through the full process.

Common Terminology Services (CTS2)

For the past several years, Mayo Clinic has been committed to the development of Common Terminology Services 2 (CTS2) through the Health Level 7 (HL7) and Object Management Group (OMG) Healthcare Services Specification Project (HSSP) collaborative process.

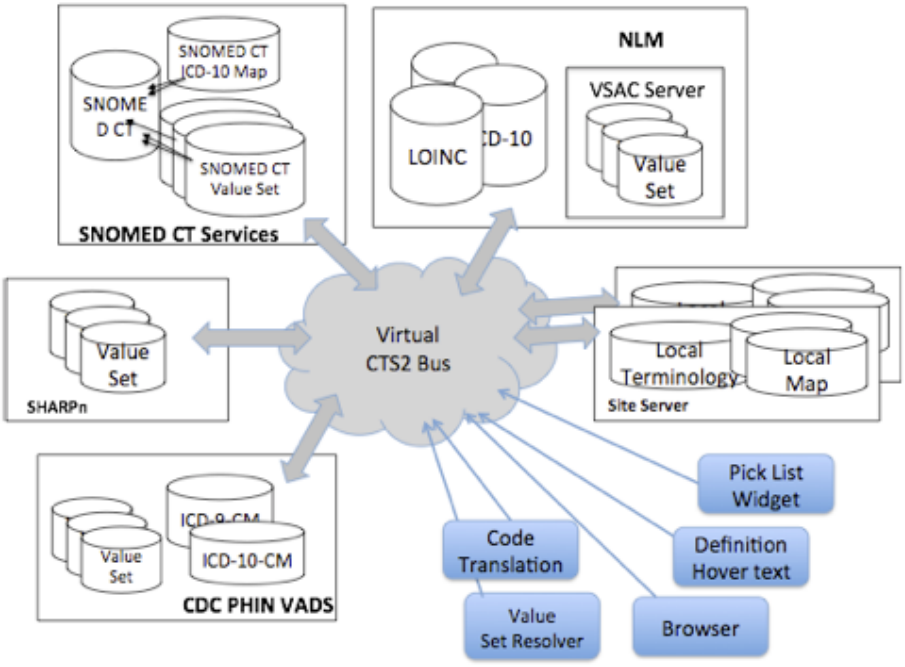
The CTS2 specification is designed to address a broad range of requirements within the ontology and terminology community. The use cases range from a need to be able to publish simple catalogs that identify what resources are available to the ability to serve the content of multiple formal ontologies, performing online reasoning and classification. The CTS2 specification also recognizes that terminological services will not necessarily be

centralized – one organization may publish catalogs, a second content and yet another may serve value sets, maps and other resources based on these tools.

The goal of this specification includes the ability to provide distributed, federated terminology services, enabling replicated service instances that periodically synchronize their content as well as service instances that reference the content in other instances. Our goal in no small part is to provide the core infrastructure that allows terminology services to be coupled and interlinked in much the same way that pages are interlinked today in the World Wide Web. Many of the design decisions that went into this specification reflect this need.

The CTS2 specification is based on the RESTful Architectural Style as described by Ray Fielding. It identifies a number of relatively fine-grained resources that have persistent identity and then describes how these resources are accessed through generic create (PUT), read (GET), update(POST or PUT) and delete(REMOVE) operations. The specification adheres to the idempotency rules laid out in Fielding's document, while introducing an additional notion of a transactional (ChangeSet) layer that allows the synchronization and exchange of collections of changes between service instances.

CTS2 allows information about value sets, maps, code systems and their corresponding content to be maintained and published anywhere on the web and, like HTML pages, to be referenced by "href" URL's. CTS2 makes it possible for value sets published by disparate organizations such as the NLM, ONC, SHARPn, IHTSDO, and others to be accessed using common tooling. It also allows maps between different code systems to be created, distributed and queried using shared tooling, much like the HTML standard enabled the development of sophisticated browsers and related tools.



As shown in the diagram on the previous page, the CTS2 standard allows multiple servers in multiple locations to be represented as a single virtual "web" of resources. Resources such as value sets and maps reference their members via URL's – URL's that can point to any node in the CTS2 web.

The Mayo team used the LexEVS service as a baseline for the Object Management Group (OMG) CTS2 submission. The underlying LexEVS model was refactored to support a Resource Oriented Architecture (ROA), which was key to creating a standard that allowed distribution and federation of terminological components. The LexEVS functional model was then used as a baseline for producing the RESTful equivalents in the CTS2 specification. The resulting standard, was finalized in August, 2012, and has been deployed in a variety of terminology service environments, including the NCBO BioPortal, the CDC PHIN VADS, the NQF eHealth initiative and others.

The following companies and organizations have indicated their support of the OMG specification: 3M Health Information Systems, Inc.; Apelon, Inc.; Everware-CBDI; Hewlett-Packard Company; Intermountain Healthcare; International Health Terminology Standards Development Organisation (IHTSDO); Model Driven Solutions; National Cancer Institute (NCI) Enterprise Vocabulary Services; NoMagic, Inc.; Sandpiper Software, Inc.; Sparx Systems; Tethers End; University of Oxford, UK, Department of Computer Science; Visumpoint.

The specification is published by the OMG and is available at http://www.omg.org/spec/cts2/1.0/ with additional information found at http://informatics.mayo.edu/cts2.

Integration of CTS2 Services with SHARPn HTP PhenotypePortal

To manage the value sets for the algorithm criteria Common Terminology Services 2 (CTS2) is being leveraged. Integrating CTS2 into PhenotypePortal allows for a single value set repository where value set versions can be managed on a per user/group basis. Since the PhenotypePortal Value Set Editor is CTS2 compliant value sets and entities can be retrieved from any other CTS2 compliant service. The value sets are initially loaded into the local repository when an algorithm is uploaded to PhenotypePortal and they are dynamically displayed when a user views an algorithms criteria. The SHARPn team plans to allow the user to modify the criteria's value sets and then execute the algorithm with their changes to see how the results are affected. Currently, the value sets and their entities for the algorithm criteria are dynamically loaded into the UI.

In Progress

- Value sets can be modified by removing entities or by adding entities from external CTS2 compliant repositories.
- Algorithms can be executed with the modified criteria (value sets).

Data Quality (DQ)

Accurate and representative EHR data are required for effective secondary use in health improvement measures and research. The variation in primary use of the data, completeness and consistency raises the question of how best to obtain information in the context of a particular secondary use of EHR data and how to provide uniform ascertainment of this information across institutions. There are often multiple sources with the same information in the EHR, such as medications from an application that supports the printing or messaging of prescriptions versus medication data that was processed from progress notes using NLP. Identical data elements from hospital, ambulatory and nursing home care, for example, have variations in their original use context. Secondary use cases may require sensitivity over specificity or vice versa, which affects the choice of source data. The metadata or knowledge base to identify all appropriate sources or interpret the source EHR's original meaning may not be available to secondary users. The SHARPn research efforts include studies of heterogeneity, quality and representativeness of *source* EHR data, in order to identify the scope, detail and source data requirements of data normalization pipeline applications.

DQ Aims:
The Data Quality team aims to develop methods and components to monitor data completeness, accuracy and suitability for the intended usage of the SHARPn deliverables. The Evaluation team is covering validation of correct functionality of the SHARPn components and services, in correctly transforming and normalizing source data. The Phenotyping team have developed user views of aggregate demographic data and patient characteristics relevant to specific phenotyping algorithms. The DQ team have focused on researching the variation across health care delivery organizations in the way medical information is recorded and coded. The DQ team also evaluate the accuracy, completeness and representativeness of source EHR data among our consortium to ascertain methods and design components to enable SHARPn pipeline users to view and assess the quality of their data with respect to their specific usage goals.

Progress:

Study 1: Studies of EHR source data heterogeneity among SHARPn collaborators
One of the SHARP use cases for automated determination of phenotype for subjects within a health system is that of the phenotype "Type 2 Diabetes Mellitus" (T2DM). Northwestern University, as a member of the eMERGE consortium, developed an algorithm for identifying cases with this phenotype based on routine EHR documentation. It would be important to understand to what extent the application of this algorithm, or any other algorithm, yields comparable case definition across Mayo Clinic and Intermountain Healthcare.

At Mayo Clinic, we have analyzed multiple sources of diagnostic information, namely the "Decision Support System" (DSS) (administrative billing data) versus the problem list in the clinical notes. These have substantial agreement as well as substantial disagreement regarding the presence or absence of Type 2 DM. Specifically, the DSS system data is more

sensitive. We are currently carrying out a chart review of the discordant results from this cross-source comparison, using the human chart reviewer as the definitive determination of disease status.

The frequency distributions for encounter diagnoses codes (the first three digits of ICD-9-CM codes) among patients identified as having at least one diagnosis code for diabetes mellitus were compared between Intermountain Healthcare's and Mayo Clinic's electronic health record (EHR). The frequencies of codes for general disease categories that may be expected to present in a primary care setting were very similar. Examples of these are

	Mayo	Intermountain
Disorders of lipoid metabolism (ICD 272.x)	73.4	76.9
Disorders of joints, NOS (ICD 719.x)	30.3	27.1
Respiratory & chest symptoms (ICD 786.x)	34.4	36.3

However, frequencies of codes for diseases and conditions seen by medical specialties, ICD 'V' codes, and codes for symptom groups tended to be dissimilar. These results will be further confirmed and evaluated for applicability as preferred ICD codes to be used in phenotyping algorithms.

Selection criteria used in the initial heterogeneity cohort were suspicious for bias in patient representation in an analysis of approximately 54,944 patients with one ICD9 code for diabetes mellitus at Intermountain Healthcare. Of these, 46,394 patients were selected according to a cohort definition designed to be compatible with a BEACON project related study. Using the constraint that patients should have at least two face to face provider encounters (as indicated by a set of E&M related CPT codes), selected cases tended to be older, have better insurance coverage and have more health care visits. We have dropped the requirement for any particular type of health care encounter in our revised cohort definition, although we will collect a summarized BETOS code status for each visit. This will enable us to evaluate potential patient selection bias according to types of visits.

Further progress on heterogeneity consists of the formation and populating of a template for delineating and comparing the various sources of data at each of our two institutions, regarding each of the potential sources of data relevant to a phenotypic definition of type 2 diabetes, namely: 1) diagnostic codes, 2) laboratory values, and 3) use of type 2 DM medications. Having completed this template, we are poised to construct an inter-institutional dataset with institution AND source within institution as 2 levels of variables that induce heterogeneity. We have also broadened the time frame constraints to allow visits and data within a 5-year time window. This will allow us to examine the impact of time windows on the resulting phenotypic data.

Study 2: BMI quality/accuracy

The accuracy and completeness of Body Mass Index (BMI) data was analyzed on both Mayo Clinic and Intermountain Healthcare data sets. Having BMI data recorded for a patient was found to be correlated with disease severity. This bias in data recording possibly resulted in higher aggregate BMI values than would have been generated if all patients had

measurements recorded. We are continuing to study this phenomenon in a more complete data set (as described above) as well as to propose remedies to alert users when data are *not* missing at random.

Study 3: Investigation of the cost benefit of an electronic algorithm for determining high-risk type 2 DM -- The "John Henry" study

A third study on data quality, in collaboration with Centerphase, seeks to study the impact of data quality on the accuracy and cost/benefit of the use of an electronic algorithm in the identification of high-risk type 2 DM cases. All data collection phases of this study have been carried out, and a draft manuscript exists, and the study is in process of manuscript preparation. The results, in brief, are that the use of the electronic algorithm reduced the number of cases that required manual screening from 200 to 34, with no apparent loss of sensitivity. The reduction in time and cost in identifying potential cases was 15% in our setting. However, the reduction in cost would increase as the number of cases to be screened or the number of required cases increases. Furthermore, the algorithm could be re-used in order to monitor the population for change in status of existing cases of high-risk type 2 DM, or for the emergence of new cases, at a much greater ratio of benefit to cost, since the algorithm development costs would not continue to accrue.

Notable Milestones Reached in DQ:

- Submitted 3 abstracts to the AMIA SUMIT March meeting, one on each of the 3 projects listed above.
- "John Henry study" data collection complete, draft manuscript prepared
- Construction of template for comparison of data sources within/across institutions
- BMI Analysis well underway
- Based on data review, flaws found in Smoking algorithm from cTAKES, reported to NLP team.

DQ Next Steps:

Heterogeneity study: Confirm and investigate similarities and differences using a more representative study population (expanded time window, less restrictive entry criteria, inclusion of control group)

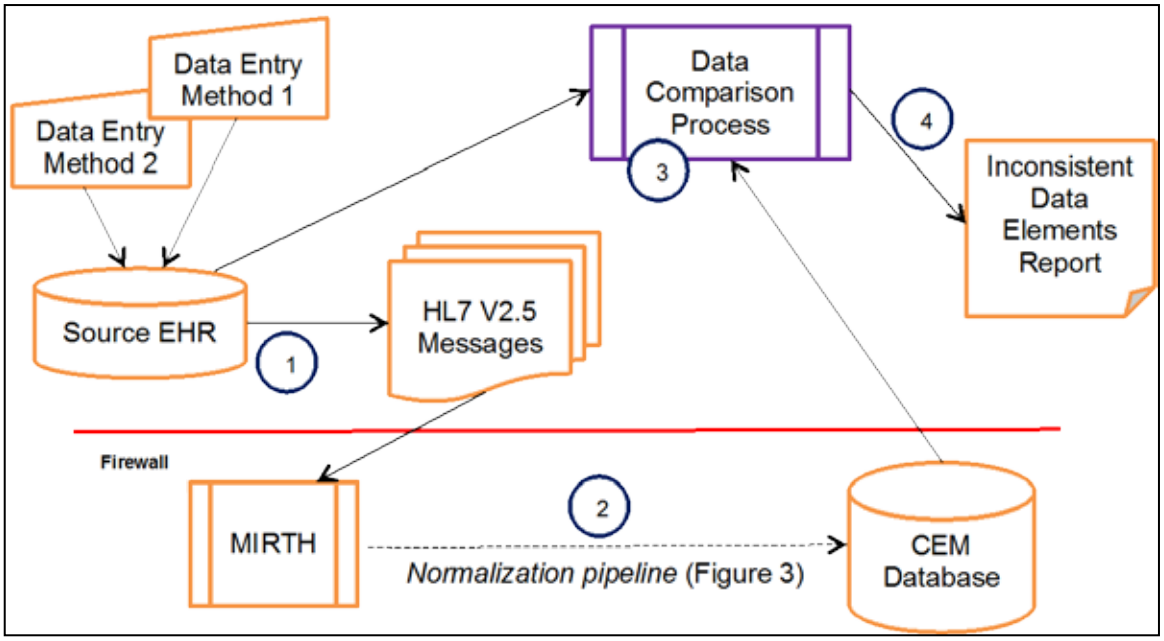
Studies of data accuracy and completeness include further analysis and characterization of the quality aspects of difficult data elements such as BMI and smoking status and develop mitigation or warnings.

Future directions

- Compare and contrast Mayo and Intermountain data
- Compare and elucidate idiosyncracies of data sources
- Draw generalization on heterogeneities
- Assess impact of these heterogeneities on secondary use

Evaluation Framework

The development of a framework within which the SHARPn data transfer and translation infrastructure can be evaluated continued through 2012 and will continue in 2013. Our approach to validating the normalization pipeline has been to construct an environment in which representative samples of various types of clinical data are transmitted across the Internet, the data are persisted as normalized CEM-based data instances, and the CEM-based instances are reconciled with the originally transmitted data. The figure illustrates this approach.



Conceptual diagram of validation processes. (1) Use case data are translated to HL7 and submitted to the MIRTH interface engine, then (2) processed and stored as normalized data objects. (3) A Java application pulls data from the source database and the CEM database, compares them, then (4) prints inconsistencies for manual review.

We have automated the majority of the process depicted here and have tested them on a stream of HL7 v2.5 medication messages. A key challenge is the translation from the terminology used by the transmitting institution to the terminology standards used within CEMs. The validation process provides functionality to identify problems in the translation processes. Our continuing objective is to develop standard procedures to identify transmission/translation errors as data from new sites, and add those data to a centralized repository. The errors expose the modifications of underlying system resources within the normalization pipeline necessary to produce an accurate conversion of the new data to the canonical CEM standard. This internal consistency is required to support the kinds of research for which this resource is intended.

Progress:

During the first months of 2012, we accomplished several tasks. These include:

Development activities produced an initial evaluation platform. This platform was based on a use case involving a simple model of medication transmittal and translation. The initial version of the platform was specifically built around the evaluation of a process that began with transmittal of HL7 2.5 medication records. The evaluation tested the accuracy of transmittal and translation of 1) the identity of the patient treated, 2) the medication administered, and 3) the time of administration.

An extension of the evaluation platform for medication transmittal and translation using the SHARPn platform was also developed. The extension of the evaluation system includes medication dose, route, and frequency. These tools used for testing have been extended so that they can be used with large numbers of records. Efforts are underway to automate as much of the process as possible.

Design and initial development of the data sets necessary for the next testing cycle of the SHARPn platform was conducted. We have been developing collections of de-identified messages appropriate for testing SHARPn software with different types of data. Our next testing cycle will be focused on laboratory data, followed by administrative, diagnostic information.

In the final months of 2012, the SHARPn Infrastructure team developed a different approach to data storage. To facilitate a collection of web-based applications, data storage was moved to Apache's CouchDB. We have redesigned the software underlying the evaluation platform and have begun the process of retesting, using the original medication records. This process will soon be complete and we will move on to further aspects of the SHARPn evaluation.

In upcoming months we intend to complete testing of all aspects of the platform for medications. In addition, we will begin testing for the next data type. We have developed a test suite for laboratory results and will soon begin the evaluation of the system with these new messages. In addition, we are designing and will soon implement a test suite using administrative data (with a focus on discharge diagnostic codes).

Milestones Reached:

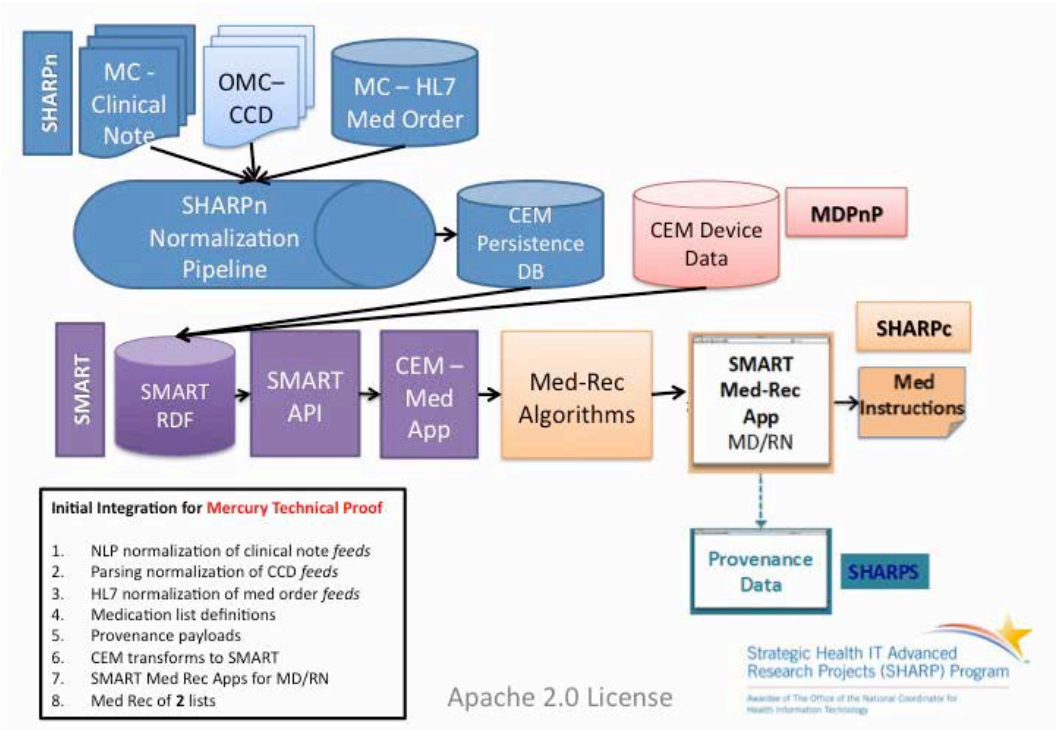
- An initial version of the evaluation platform has been developed and tested.
- The existing approach has been upgraded to accommodate the new SHARPn data storage model. Testing of this upgrade is underway.
- Several different data sets have been extracted from the Intermountain Healthcare Enterprise Data Warehouse (EDW), have been de-identified, and have been configured as streams of HL7 messages.
- Successful transmission of messages across the Internet has been demonstrated repeatedly. We have moved our focus to the development of those parts of the system that will allow us to characterize and correct those errors that occur in the data translation components of the system.

Next Steps:

- Complete testing of the translation of medication messages using the new data storage facilities.
- Begin testing different message types, including laboratory results and administrative data.
- Achieve our goal of demonstrating the evaluation platform over a range of data types. We will focus on those that will provide an example of a process that begins with the transmittal of HL7-based data from one institution to another and ends with the successful execution of a phenotyping algorithm run against this transmitted and translated data.

PAN-SHARP

A key demonstration of the tools developed in SHARPn and put to the test in a joint effort across all of the SHARP program participants. The challenge put forth in 2011, was to identify a problem where all SHARP participants could synergistically contribute. The cross-SHARP projects agreed upon Medication Reconciliation with a vision to create a novel, functional Medication Reconciliation solution for use in a clinical environment. The proof-of concept was intended to improve the state of Healthcare IT and at the same time highlight the SHARP members' contributions and technical innovations.

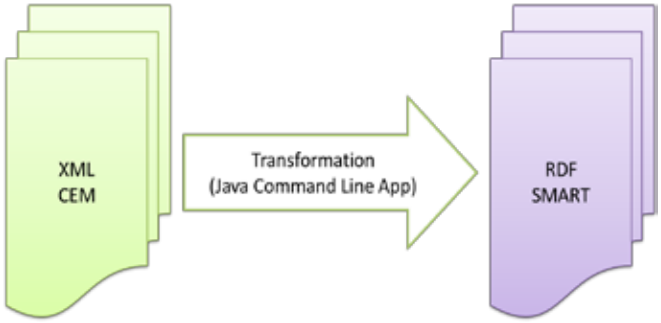


- Security of data management – SHARPs
- Data normalization for secondary use – SHARPn
- Device generated drug data – NIH Device Affiliate
- Creation of SmartApp – Smart
- Cognitive interface and refinement - SHARPc

The Pan-SHARP's project ultimate goal was to become a viable medication reconciliation solution for the real world, thus indispensable to receive data from as many sources as possible. However, in order to be able to compute on this data, it must have a single, common representation. The normalization and infrastructure teams were key in this "Pan-SHARP" effort to show how data from various sources and formats can be normalized to CEM instances and then exposed to a SMART Platform medicine reconciliation application. The team started with collaborative discussions to create the models required by the application and a transform from the CEM format to the format needed by the SMART Platform application. This was an extension of previous work in collaboration with Area 3, or the SMART Platform team on CEM models and persistent layer cohesion.

The SHARPn team partnered with the Southeast Minnesota Beacon Community to obtain real data from disparate sources to process for the Pan SHARP evaluation. Upon IRB approvals, the team leveraged the Rochester Epidemiology Project, to randomly sample 400 patients in Olmsted County where an EMR record was identified as existing in both Mayo Clinic and Olmsted Medical Center (OMC) during 2010. For those overlap patients, each site provided clinical data via multiple data sources.

	Mayo	OMC
Clinical Text	CDA	
Structured Data	HL7 V2	CCD



All of the data was processed using SHARPn data normalization pipeline.

- Normalize to NotedDrug CEMs
- De-Identify:
  - Add or subtract a random fraction of five years to the age.
  - Randomly shift all dates by some number of months, and then fuzz each day +/- 3 days.
- Compose three medication lists (CCD, HL7, Clinical Notes) for other teams

XML CEM files were delivered in a compressed zip file. These files were processed with a java based command line application written for the transformation. The resultant RDF files were zipped and delivered to the SMART team.

Program Outputs

a) Products

i) Data Normalization / CEM Database Design
   (1) Models, value sets, and pipeline for demographics, labs, drugs, procedures and disorders
   (2) Mirth channels and APIs for CouchDB-based CEM databases
   (3) CEM request/browser website
   (4) CTS2 SHARPn instance

ii) NLP
   (1) March, 2012 - negation, uncertainty, and conditional attribute annotator incorporated into cTAKES 2.5. This version has been included as part of the Apache cTAKES project
   (2) March, 2012 - Subject and Generic attribute annotator released (via the assertion module in cTAKES 2.5)
   (3) April, 2012 – cTAKES 2.0 with the implementation of the Common Type System and a new tokenizer following Penn Treebank tokenization rules
   (4) May, 2012 – cTAKES 2.5 with a new assertion module, a new part of speech model trained on more clinical data, an additional sectionizer (MIT), updates on the co-reference resolver annotator, a semantic role labeler as part of the dependency parser
   (5) May, 2012 - New, improved models were developed for part-of-speech tagging, dependency parsing, and semantic role labeling
   (6) June, 2012 – cTAKES accepted as an Incubator project into the Apache Software Foundation (ASF). This event is expected to have a wide impact on both the national and international communities as ASF is the world-renowned organization for open source projects housing such projects as HTTP server, UIMA, Hadoop, Lucene, OpenOffice, Perl, Subversion, Tomcat.
   (7) September, 2012 - Subject and Generic attribute annotator included in Apache cTAKES
   (8) October, 2012 - The fully functional relation module is available for download as part of Apache cTAKES. The released module includes the models trained on the SHARP data. This includes the models for the discovery of the bodySite and severity modifiers
   (9) October, 2012 – Improved Apache cTAKES Drug NER module
   (10) October, 2012 - Beta release of template population delivered to infrastructure team. This was built on cTAKES 2.5
   (11) December, 2012 – Apache cTAKES (incubating) 3.0.0 RC4. First release of an Apache (incubator) project.
   (12) The NLP Evaluation Workbench has been made available to interested users via a web site managed by the University of California at San Diego

iii) Infrastructure

a) Publications and Presentations (2012)

i) Recent/accepted/published

1) Albright D, Lanfranchi A, Fredriksen A, Styler W, Warner C, Hwang J, Choi J, Dligach D, Nielsen R, Martin J, Ward W, Palmer M, Savova G. Towards syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association. (2013. forthcoming).
2) Wagholikar K, Torii M, Jonnalagadda S, Liu H. Pooling annotated corpora for clinical concept extraction. Journal of Biomedical Semantics. 2013; forthcoming
3) Torii M, Wagholikar K, Liu H. Detecting concept mentions in biomedical text using Hidden Markov Model: Multiple concept types at once or one at a time. Journal of Biomedical Semantics. 2013; forthcoming
4) Sohn S. Computational Semantics in Clinical Text (CSCT) workshop, Potsdam, Germany, accepted (journal eligible). 2013; forthcoming
5) Pathak J, Bailey K, Beebe C, Carrell D, Hart L, Haug P, Huff S, Kaggal V, Li D, Liu H, Marchant K, Oniki T, Rea S, Savova G, Solbrig H, Tao C, Taylor D, Westberg L, Zhuo N, Chute C. (under review). Strategic Health IT Research Project on Normalization and Standardization of Electronic Health Record Data for High-Throughput Phenotyping: the SHARPn Consortium. Journal of Health Services Research. Special issue (14): Health Information Technology. 2013.
6) Wu ST, Kaggal VC, Dligach D, Masanz JJ, Chen P, Becker L, Chapman WW, Savova GK, Liu H, Chute CG. A common type system for clinical Natural Language Processing. Journal of Biomedical Semantics. 2013.
7) Jonnalagadda S, Cohen T, Wu S, Liu H, Gonzalez G. Evaluating the Use of Empirically Constructed Lexical Resources for Named Entity Recognition. Will appear in Computational Semantics in Clinical Text. 2013
8) Sohn S, Murphy SP, Jonnalagadda S, Wagholikar K, Halgrim S, Wu ST, Chute CG, Liu H. Systematic Analysis of Cross-Institutional Medication Description Patterns in Clinical Notes. Will appear in Computational Semantics in Clinical Text. 2013.
9) Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, Ravikumar KE, Wu ST, Kullo IJ, Chute CG. An Information Extraction Framework for Cohort Identification Using Electronic Health Records. AMIA CRI 2013. Forthcoming
10) Jiang G, Liu H, Solbrig HR, Chute CG. ADEpedia 2.0: Integration of Normalized Adverse Drug Events (ADEs) Knowledge from the UMLS. AMIA CRI 2013. Forthcoming
11) Liu H, Wu ST, Li D, Jonnalagadda S, Sohn S, Wagholikar K, Haug PJ, Huff SM, Chute CG. Towards a semantic lexicon for clinical natural language processing. AMIA. 2012.
12) Sohn S. "Systematic Analysis of Cross-Institutional Medication Description Patterns in Clinical Notes" IEEE HISB 2012 (poster)
13) Miller T, Dligach D, Savova G. Active learning for Coreference Resolution in the Biomedical Domain. BioNLP workshop at the Conference of the North American Association of Computational Linguistics (NAACL 2012). Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pp. 73-81. 2012.

14) Pestian J, Deleger L, Savova G, Dexheimer J, Solti I. Natural Language Processing, the basics. In Pediatric Biomedical Informatics: Computer Application in Pediatric Research (Translational Bioinformatics). Ed. John J. Hutton. Springer, 2012. ISBN-10: 9400751486, ISBN-13: 978-9400751484.
15) Savova G, Deleger L, Solti I, Pestian J, Dexheimer J. Natural Language Processing: Applications in Pediatric Research. In Pediatric Biomedical Informatics: Computer Application in Pediatric Research (Translational Bioinformatics). Ed. John J. Hutton. Springer, 2012. ISBN-10: 9400751486, ISBN-13: 978-9400751484.
16) Jonnalagadda SR, Li D, Sohn S, Wu ST, Wagholikar K, Torii M, Liu H. Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules. J Am Med Inform Assoc. 2012 Sep 1; 19(5):867-74. Epub 2012 Jun 16. PMID:22707745. PMCID:3422831. DOI:10.1136/amiajnl-2011-000766.
17) Wagholikar KB, MacLaughlin KL, Henry MR, Greenes RA, Hankey RA, Liu H, Chaudhry R. Clinical decision support with automated text processing for cervical cancer screening. J Am Med Inform Assoc. 2012 Sep 1; 19(5):833-9. Epub 2012 Apr 29. PMID:22542812. PMCID:3422840. DOI:10.1136/amiajnl-2012-000820.
18) Liu H, Wu S, Tao C, Chute C. Modeling UIMA Type System Using Web Ontology Language – towards Interoperability among UIMA-based NLP Tools. MIXHS. 2012.
19) Wagholikar K, Sohn S, Wu S, Kaggal V, Buehler S, Greenes R, Wu T, Larson D, Liu H, Chaudhry R, Boardman L. Clinical Decision Support for Colonoscopy Surveillance Using Natural Language Processing. The Second IEEE Conference on Healthcare Informatics, Imaging, and Systems Biology (HISB). 2012.
20) Jonnalagadda SR, Del Fiol G, Medlin R, Weir C, Fiszman M, Mostafa J, Liu H. Automatically extracting sentences from Medline citations to support clinicians' information needs. J Am Med Inform Assoc. 2012 Oct 25. [Epub ahead of print]
21) Li DC, Endle CM, Murthy S, Stancl C, Suesse D, Sottara D, Huff SM, Chute CG, Pathak J. Modeling and Executing Electronic Health Records Driven Phenotyping Algorithms using the NQF Quality Data Model and JBoss Drools Engine. American Medical Informatics Association (AMIA) Annual Symposium Proceedings. 2012. (Epub ahead of print).
22) Pathak J, Kiefer R, Freimuth R, Chute C. Validation and discovery of genotype-phenotype associations in chronic diseases using linked data. Stud Health Technol Inform. 2012; 180:549-53. PMID:22874251.
23) Pathak J, Kiefer RC, Bielinski SJ, Chute CG. Mining the Human Phenome using Semantic Web Technologies: A Case Study for Type 2 Diabetes. American Medical Informatics Association (AMIA) Annual Symposium Proceedings. 2012 (Epub ahead of print).
24) Pathak J, Kiefer RC, Bielinski SJ, Chute CG. Applying Semantic Web Technologies for Phenome-Wide Scan using an Electronic Health Record Linked Biobank Journal of Biomedical Semantics (JBSM). 2012 (Epub ahead of print).
25) Pathak J, Kiefer RC, Chute CG. Applying Linked Data principles to represent patient's electronic health records at Mayo Clinic: A case report. IHI'12 -

Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. 2012:455-64.

26) Pathak J, Kiefer RC, Chute CG. Using semantic web technologies for cohort identification from electronic health records to conduct genomic studies. AMIA Summits Transl Sci Proc. 2012; 2012:10-9. Epub 2012 Mar 19. PMID:22779040. PMCID:3392057.
27) Thompson WK, Rasmussen LV, Pacheco JA, Peissig PL, Denny JC, Kho AN, Miller A, Pathak J. An Evaluation of the NQF Quality Data Model for Representing Electronic Health Record Driven Phenotyping Algorithms. American Medical Informatics Association (AMIA) Annual Symposium Proceedings. 2012 (Epub ahead of print).
28) Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, Waudby C, Chen L, Denny JC, Wilke RA, Pathak J, Carrell D, Kho AN, Starren JB. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. J Am Med Inform Assoc. 2012 Mar-Apr; 19(2):225-34. PMID:22319176. PMCID:3277618. DOI:10.1136/amiajnl-2011-000456.
29) Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, Tao C, Parker CG, Haug PJ, Huff SM, Chute CG. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. J Biomed Inform. 2012 Aug; 45(4):763-71. Epub 2012 Feb 04. PMID:22326800. DOI:10.1016/j.jbi.2012.01.009.
30) Clifford L, Singh A, Wilson GA, Toy P, Gajic O, Malinchoc M, Herasevich V, Pathak J, Kor DJ. Electronic health record surveillance algorithms facilitate the detection of transfusion-related pulmonary complications. Transfusion. 2012 Aug 31. [Epub ahead of print] PMID:22934792. DOI:10.1111/j.1537-2995.2012.03886.x.
31) Tao C, Pathak J, Solbrig HR, Wei WQ, Chute CG. Terminology representation guidelines for biomedical ontologies in the semantic web notations. J Biomed Inform. 2012 Sep 28. [Epub ahead of print] PMID:23026232. DOI:10.1016/j.jbi.2012.09.003.
32) Tao C, Jiang G, Oniki TA, Freimuth RR, Zhu Q, Sharma D, Pathak J, Huff SM, Chute CG. A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. J Am Med Inform Assoc. 2012 Dec 25. [Epub ahead of print] PMID:23268487. DOI:10.1136/amiajnl-2012-001326.
33) Bielinski S, Pathak J, Liu H, Sohn S, Jarvik G, Carrell D, Pereira N, Roger V. Using Electronic Health Records to Identify Heart Failure Cohorts with Differentiation for Preserved and Reduced Ejection Fraction (Poster). AMIA 2012 Annual Symposium. Nov 2012.
34) Endle C, Murthy S, Suesse D, Stancl C, Li DC, Hart L, Chute CG, Pathak J. Visualization and Reporting of Results for Electronic Health Records Driven Phenotyping using the Open-Source popHealth Platform (poster). AMIA Annual Symposium. Nov 2012.
35) Li DC, Shrestha G, Murthy S, Sottara D, Huff SM, Chute CG, Pathak J. Applying JBoss® Drools Business Rules Management System for Electronic Health Records Driven Phenotyping. AMIA Annual Symposium 2012. Nov 2012.

36) Pathak J, Al-Kali A, Talwalkar JA, Kho AN, Denny JC, Murphy SP, Bruce KT, Durski MJ, Chute CG. Using Electronic Health Records to Identify Patient Cohorts for Drug-Induced Thrombocytopenia, Neutropenia and Liver Injury. AMIA Annual Symposium. Nov 2012.
37) Pathak J, Kiefer RC, Freimuth RR, Bielinski SJ, Chute CG. Mining Genotype-Phenotype Associations from Electronic Health Records and Biorepositories using Semantic Web Technologies (poster). AMIA Annual Symposium. Nov 2012.
38) Teagno J, Kiefer R, Pathak J, Zhang GQ, Sahoo S. A Distributed Semantic Web Approach for Cohort Identification (Poster). AMIA 2012 Annual Symposium. Nov 2012.
39) Jiang G, Solbrig HR, Chute CG. Using semantic web technology to support ICD-11 textual definitions authoring. ACM International Conference Proceeding Series. 2012; 38-44.
40) Pathak J, Kiefer RC, Chute CG. The Linked Clinical Data project: Applying Semantic Web technologies for clinical and translational research using electronic medical records. ACM International Conference Proceeding Series. 2012; 94-5.
41) Pathak J, Weiss LC, Durski MJ, Zhu Q, Freimuth RR, Chute CG. Integrating va's ndf-rt drug terminology with pharmgkb: preliminary results. Pac Symp Biocomput. 2012; 400-9. PMID:22174295.
42) Tao C, Wongsuphasawat K, Clark K, Plaisant C, Shneiderman B, Chute CG. Towards event sequence representation, reasoning and visualization for EHR data. IHI'12 - Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. 2012:801-5.
43) Sohn S, Wu ST, Chute CG. Dependency Parser-based Negation Detection in Clinical Narratives. AMIA Summit on Clinical Research Informatics (CRI). San Francisco, CA. Mar 2012.
44) Song D, Chute CG, Tao C. Semantator: Annotating Clinical Narratives with Semantic Web Ontologies AMIA Clinical Research Informatics. Mar 2012.
45) Wu ST, Liu H, Li D, Tao C, Musen MA, Chute CG, Shah NH. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. J Am Med Inform Assoc. 2012 Jun 1; 19(e1):e149-56. Epub 2012 Apr 04. PMID:22493050. DOI:10.1136/amiajnl-2011-000744.
46) Sohn S, Torii M, Li D, Wagholikar K, Wu S, Liu H. A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes Biomedical Informatics Insights. 2012(5 Suppl 1):43-50.
47) Wagholikar K, Torii M, Jonnalagadda S, Liu H. Feasibility of pooling annotated corpora for clinical concept extraction. AMIA Summit on Clinical Research Informatics (CRI). San Francisco, CA. 2012.
48) Liu H, Wagholikar K, Wu S. Using SNOMED CT to encode summary level data - a corpus analysis. AMIA Summit on Clinical Research Informatics (CRI). San Francisco, CA. Mar 2012.
49) Liu M, Shah A, Jiang M, Peterson N, Dai Q, Aldrich M, Chen Q, Bowton E, Liu H, Denny J, Xu H A Study of Transportability of an Existing Smoking Status Detection Module across Institutions. 2012 AMIA Symposium. Nov 2012.

50) Chasin R, Rumshisky A, Uzuner Ö, Szolovits P. Using UMLS for Word Sense Disambiguation in Clinical Notes. AMIA 2012. Chicago, IL.
51) Jonnalagadda S, Del Fiol G, Medlin R, Weir C, Fiszman M, Mostafa J, Liu H. Automatically Extracting Sentences from Medline Citations to Support Clinicians' Information Needs. Journal of American Medical Informatics Association. 2012; doi:10.1136/amiajnl-2012-001347 [Online First]

i) Planned
1) Journal submission describing the full gold standard created under SHARP. The syntactic and UMLS entities have already been described in Albright et al, 2013.
2) Journal submission describing methods and evaluations on the SHARP Seed Corpus
3) Journal submission describing the results produced by different entity disambiguation, and evaluation using additional corpora
4) Journal submission describing methods and evaluations for negation and uncertainty on the SHARP Seed Corpus
5) Journal manuscript describing the methods for discovering subject and generic attributes
6) Journal submission describing the relation extraction methods including the discovery of bodySite and severity attributes. It will provide the details of the implementations, the contributions of different types of features used by the module, and the results of our experiments with tree kernels.
7) Journal submission about the MedER system.
8) Journal submission "Analysis of Cross-Institutional Medication Information Annotations in Clinical Notes"
9) Journal submission on improved coreference system on SHARP data and integration with other systems such as relation extraction
10) AMIA 2013 submission on the new machine learning sectionizer

Events

January 12- 14, 2012: Clinical Information Modeling Initiative (CIMI), San Antonio, TX
CIMI's 5th group meeting was held. Over 35 people attended in person with representation from SHARPn PI's Drs. Stan Huff and Christopher Chute.

January 27, 2012: PanSHARP MedRed In-Person Kickoff, Washington, DC
SHARPn represented by Lacey Hart, Calvin Beebe, Tom Oniki.

February 20-24, 2012: Healthcare Information and Management Systems Society (HIMSS), Las Vegas, NV
SHARPn tools in the Exhibition with other SHARP and ONC programs.

February 22-24, 2012: Conference on Semantics in Healthcare and Life Sciences (CSHALS), Cambridge/Boston, MA
Representing SHARPn was Dr. Jyoti Pathak

March 19-23, 2012: American Medical Informatics Association Clinical Research Informatics (AMIA CRI), San Francisco, CA
SHARPn members in attendance presented and participated in the working sessions.

April 4-5, 2012: NQF Annual Conference, Washington, DC
Representing SHARPn was Dr. Jyoti Pathak

April 11-13, 2012: ONC Standards and Interoperability (S&I) Framework Face-to-Face (F2F) Conference, Alexandria, VA
SHARPn members in attendance participated in the working sessions.

April 23-24, 2012: NIH Workshop: NLP CDS, Bethesda, MA
SHARPn members in attendance participated in the working sessions.

April 25-26, 2012: CTSA Clinical and Translational Science Ontology Workshop, Baltimore, MD
Representing SHARPn was Dr. Jyoti Pathaki

May 10-12, 2012: Clinical Information Modeling Initiative (CIMI), Pleasanton, CA
Representing SHARPn was Dr. Stan Huff and Harold Solbrig.

May 31, 2012: National Cancer Informatics (NCI) Meeting, Bethesda, MD
Dr. Christopher Chute participated in an organized discussion on NCIP.

June 2012
Applying dependency parses and SRL: Subject and Generic Attribute Discovery.
Presentation by Stephen Wu.

June 2012
Centers for Translational Science Activities (CTSA) Toolshop Webinar "Clinical Text Analysis and Knowledge Extraction System (cTAKES)".
Webinar to a national CTSA audience by Pei Chen, Pei, Hongfang Liu; James Masanz, and Guergana Savova.

June 11-12, 2012: Area 4 SHARP Face to Face Conference; Rochester MN
125 Attendees with representatives from SHARPn, other SHARP programs, ONC, FSC, PAC, and Beacons.

June 14, 2012: 8th Annual MN e-Health Summit, Brooklyn Park, MN
Dr Christopher Chute and Dr. Jyoti Pathak presented "The SHARPn Project on Secondary Use of Electronic Medical Record Data".

June 23-26, 2012: 2012 Academy Health Annual Research Meeting, Orlando, FL
SHARPn members in attendance participated in the working sessions.

July 21-25, 2012: ICBO 2012, Graz, Austria
Dr. Guoqian Jiang represented SHARPn

August 23, 2012: IOM Roundtable: Digital Learning Collaborative, Washington, DC
Dr. Christopher Chute represented SHARPn.

September, 2012
Active learning for coreference. Natural Language Processing (NLP) Annotation workshop collocated with the 2nd annual IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology, San Diego, CA.

September, 2012
Natural Language Processing (NLP) Annotation workshop collocated with the 2nd annual IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology "Shared Annotated Resources for the Clinical Domain", San Diego, CA, USA. Presentation by Guergana Savova.

September 10-11, 2012 Harvard SHARP F2F, Cambridge, MA
Collaboration meeting attended by SHARPn members.

September 14-16, 2012: Clinical Information Modeling Initiative (CIMI), Rockville, Maryland
Dr. Stan Huff and Harold Solbrig represented SHARPn.

September 21, 2012: NQF: Advancing Solutions for eMeasure Implementation, Washington DC
SHARPn members in attendance participated in the working sessions.

September 28, 2012: IEEE Conference at HISB 2012, San Diego, CA
Dr. Christopher Chute represented SHARPn.

October, 2012
Natural Language Processing Working Group Pre-Symposium – doctoral consortium and a data workshop "Shared Annotated Resources for the Clinical Domain". American Medical Informatics Association. Presentation by Guergana Savova.

October, 2012
Annual Centers for Translational Science Activities (CTSA) meeting "SHAPPn: Clinical Natural Language Processing". Chicago, IL, USA. Presentation by Guergana Savova.

October 1-3, 2012: Computational Individualized Medicine Workshop, 2012, Rochester, MN
SHARPn members in attendance participated in the working sessions.

October 22, 2012: IntelliFest 2012, San Diego, CA
Dr. Jyoti Pathak represented SHARPn

October 28-November 2, 2012: Managing Interoperability and complexity in Health Systems (MIX-HS) / 21st ACM International Conference on Information and Knowledge Management (CIKM), Maui, HI
Drs. Cui Tao and Guoqian Jiang represented SHARPn

November 1-3, 2012: 2012 International Symposium on Big Data and MapReduce, Xiangtan Hunan, China
Dr. Dingcheng Li represented SHARPn

November 3-7, 2012: American Medical Informatics Association (AMIA), Chicago, IL
SHARPn attendees were involved in 44 different sessions involving panels, demonstrations, paper, presentations and poster

November 7, 2012: CTSA Annual Informatics Meeting, Chicago, IL
Representing SHARPn, Dr. Dingcheng Li, Dr. Hongfang Liu, Dr. Cui Tao. SHARPn also lead a breakout session.

November 13-15, 2012: NLM CTS2 Meeting, Bethesda, MD
SHARPn members in attendance participated in the working sessions.

December 2-4, 2012: CIMI Modeling Face to Face, Amsterdam, Netherlands
SHARPn represented by Stan Huff and Harold Solbrig

December 11, 2012: SHARPFest; Washington, DC
Attendees: Christopher Chute, Stan Huff, Lacey Hart, Jyoti Pathak, Guergana Savova, Tom Oniki

December 11-13, 2012: All ONC Grantee Meeting, Washington DC
Attendees: Christopher Chute, Lacey Hart, Jyoti Pathak.

Partnerships / Relationships / Alliances (New/Ongoing in 2012)

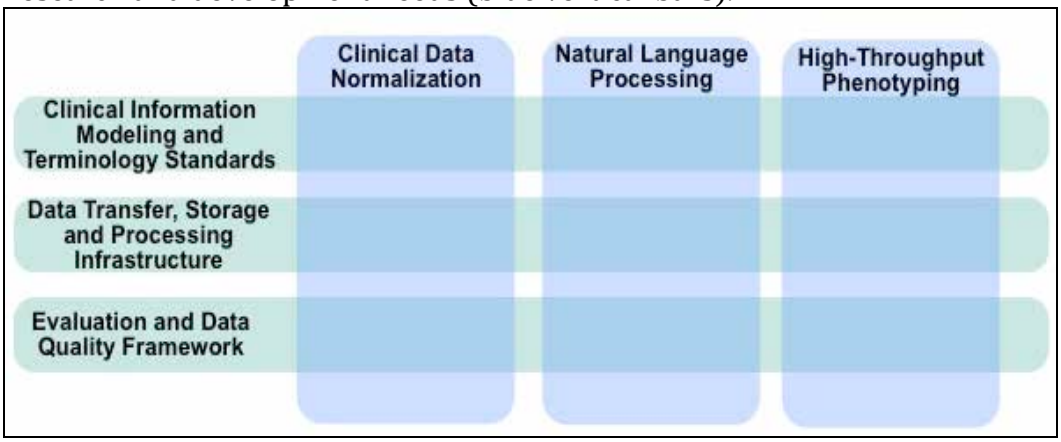
1) Apache Software Foundation
2) Center for Integration of Medicine & Innovative Technology (CIMIT); MD SHARP project - Medical Device Plug-and-Play (MD PnP) Interoperability Program
3) Clinical Information Modeling Initiative (CIMI)- International consensus group with Detailed Clinical Models
4) Consortium for Healthcare Informatics Research (CHIR)
5) Informatics for Integrating Biology and the Bedside (i2b2)
6) Electronic Medical Records and Genomics (eMERGE) Network
7) Multi-source Integrated Platform for Answering Clinical Questions (MiPACQ)
8) National Center for Cognitive Informatics and Decision Making in Healthcare (NCCD); SHARP Project (SHARPc) - Cognitive Foundations for Decision Making
9) National Library of Medicine, UMLS team
10) National Quality Forum / MAT User Group
11) Open Health Natural Language Processing (OHNLP)
12) Pharmacogenomics Research Network (PGRN) – PGPop team
13) Southeast Minnesota Beacon Community
14) Standards & Interoperability (S&I) Framework - Query Health
15) Substitutable Medical Apps, reusable technologies (SMArt); Harvard SHARP Team
16) Temporal Histories for Your Medical Events (THYME; R01 LM 10090 funded by NLM)
17) University California, San Diego National Center for Biomedical Computing (NCBC) Integrating Data for Analysis, Anonymization and SHaring (iDASH)
18) University of Illinois at Urbana-Champaign; SHARP on Security (SHARPs)

Operational Activities

Collaboration has continued to be fostered in both intra-SHARP and cross-SHARP programs in cross-knowledge pollination and collaboration activities. SHARPn provided all of the project management support for the Pan-SHARP project.

Mayo Clinic serves as the coordination hub for all SHARPn activities, providing a high standard of accountability in achieving program goals. The program is under the direction of experienced, qualified, Mayo Clinic Project Management Institute (PMI)-certified project managers. Project managers are responsible for day-to-day management, execution, and delivery of project team deliverables. The project managers track progress (scope, resources & costs), proactively manage risk, track lessons learned & report to stakeholders.

In the past two years, we have developed and maintained an organizational and functional framework for SHARPn that is structured around infrastructure (green horizontal bars) as well as research and development needs (blue vertical bars).



The SHARPn program held its 3rd annual meeting on **June 11-12, 2012** on the University of Minnesota Rochester Center campus to present and discuss work in the following areas:

- Standards, Data Integration & Semantic Interoperability
- Natural Language Processing
- Phenotyping: clinical trial selection, clinical decision support, quality measurement
- Software evaluation

Over 125 attendees with representatives from SHARPn, other SHARP programs, and the ONC were present. Attendees:

- **Learned** more about SHARPn and the health IT field
- **Explored** technologies and tools that enable secondary uses of EHR data
- **Participated** in a hands-on tutorial from experts on how to better leverage secondary data using open source tools
- **Presented** papers or posters on related topics through a peer reviewed process
- **Worked side-by-side** with community leaders to develop and refine customer and stakeholder requirements

1) Personnel / Hiring (ARRA Report)

Budgeted Personnel have remained consistent to projections with approved justification approved.

Calendar Year / Quarter:	2011 / 1	Calendar Year / Quarter:	2012/ 1
Number of Jobs Count:	14.3	Number of Jobs Count:	19.06
Calendar Year / Quarter:	2011 / 2	Calendar Year / Quarter:	2012 / 2
Number of Jobs Count:	13.5	Number of Jobs Count:	19.06
Calendar Year / Quarter:	2011 / 3	Calendar Year / Quarter:	2012 / 3
Number of Jobs Count:	13.5	Number of Jobs Count:	18.60
Calendar Year / Quarter:	2011 / 4	Calendar Year / Quarter:	2012 / 4
Number of Jobs Count:	22.44	Number of Jobs Count:	18.60

2) Grants Management (ARRA Report)

As a matrixed program, project expenditures are spread across the consortium. Expenditures have remained consistent with work scope approved; noted changes from the original notice of award reflect ….

Calendar Year / Quarter:	2012 / 1
Total Federal Amount of ARRA Expenditure:	$6,009,713
Calendar Year / Quarter:	2012 / 2
Total Federal Amount of ARRA Expenditure:	$7,357,340
Calendar Year / Quarter:	2012 / 3
Total Federal Amount of ARRA Expenditure:	$8,967,985

Glossary

API: Application Programming Interface
BEACON: Beacon Communities Program, ONC Funded
CCD: Continuity of Care Document
CDA: Clinical Document Architecture
CCDA: Consolidated Clinical Document Architecture
CEM: Clinical Element Model
CHIR: Consortium for Healthcare Informatics Research
CIMI: Clinical Information Modeling Initiative
CIMIT: Center for Integration of Medicine & Innovative Technology
CTS2: Common Terminology Services 2
EHR: Electronic Health Record
eMERGE: Electronic Medical Records and Genomics Network
EMR: Electronic Medical Record
HIPAA: Health Insurance Portability and Accountability Act of 1996
HL7: Health Level Seven International
HSSP: Healthcare Services Specification Project
i2b2: Informatics for Integrating Biology and the Bedside
iDASH: Integrating Data for Analysis, Anonymization and SHaring
IE: Information Extraction
IHC: Intermountain Health Care
MD SHARP: NIH SHARP Program Affiliate
MiPACQ: Multi-source Integrated Platform for Answering Clinical Questions
ML: Machine Learning
MU: Meaningful Use
NLM: National Library of Medicine
NLP: Natural Language Processing
NQF: National Quality Forum
NwHIN: Nationwide Health Information Network
OHNLP: Open Health Natural Language Processing
OMG: Object Management Group
ONC: Office of the National Coordinator for Health Information Technology
OWL: Web Ontology Language
QDM: Quality Data Model
PGRN: Pharmacogenomics Research Network
RDF: Resource Description Framework
REST: Representational State Transfer
SEMNBC: Southeast Minnesota Beacon Community
SHARP: Strategic Health IT Advanced Research Projects, ONC funded
SHARPc: SHARP Area 2 Projects on patient centered cognitive support research
SHARPn: SHARP Area 4 Projects on Secondary Use of EHR Data Program
SHARPs: SHARP Area 1 Projects on Security
S&I: Standards & Interoperability Framework, ONC Funded
SMArt: SHARP Area 3 Substitutable Medical Apps, reusable technologies
SWRL: Semantic Web Rule Language
THYME: Temporal Histories for Your Medical Events
UIMA: Unstructured Information Management Architecture
UMLS: Unified Medical Language System