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: 1/1/2011 – 12/31/2011

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®

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 (EHR). 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 assembled 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.

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

In order to represent detailed clinical data models, the Clinical Element Model (CEM) originating from Intermountain Healthcare has been leveraged.

DN Aims: Build a generalizable data normalization pipeline, establish a globally available resource for health terminologies and value sets, and establish and expand modular library of normalization algorithms. Iteratively test normalization pipelines, including textual normalization where appropriate, against normalized forms, and tabulate discordance. Use cohort identification algorithms in both Electronic Medical Record (EMR) data and data warehouse data.

Progress: The Data Norm team has worked extensively with clinical model needs from a variety of perspectives. For the SHARPn and "Pan-SHARP" needs, they have created models and terminology that will be used as the normalization targets in the areas of demographics, labs, drugs, and disorders. We followed a strategy of creating "core" models that will contain a superset of attributes needed by various contexts (including the SHARPn secondary use context) and constraints on those models useful to the secondary use context.

In an international context to addressing clinical models, Dr. Stan Huff has introduced an international consensus group with detailed clinical models, Clinical Information Modeling Initiative (CIMI). This group is tackling the following:

i) Archetype Object Model/ADL 1.5 openEHR
ii) CEN/ISO 13606 AOM ADL 1.4
iii) UML 2.x + OCL + health care extensions

iv) OWL 2.0 + health care profiles and extensions
v) MIF 2 + tools HL7 RIM – static model designer

A website has been built (in development) from which new models and terminology or changes to existing models and terminology can be requested. The website requires the user to create a simple logon account, after which he/she may create new requests, view requests made by his/her team, and modify/add information to requests he/she has made. This website will facilitate the communication process and allow the Intermountain Healthcare modeling team to capture all requests in a single database.

Also in development is a website from which a user may search/browse model content. The website will present a graphical ("tree") view of a model as well as XML/CDL views. From the website, the user will be able to download model files and terminology. It will contain a link to the CEM request website.

A subgroup focuses on the review the CDISC standards published and discusses approach for the alignment with the SHARP core CEM models. Initial focus of the group was on the domains of laboratory, medication and demographics and to standardize the core CEMs with existing data standards (e.g. CDISC standards, ISO 11179 standard).

The data normalization team is also key in the "Pan-SHARP" effort in which the various SHARP grantees will come together 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 has begun 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 is an extension of previous work in collaboration with Area 3, or the SMART Platform team on CEM models and persistent layer cohesion.

Notable Milestones Reached in DN:

1) In collaboration with the Regenstrief Institute, this team decomposed the Regenstrief HOSS Pipeline into SHARP 4 computing environment and mapped its move into UIMA platform
2) Design of other components of data normalization framework (terminology services - NwHIN connections)
3) Completion of 'Tracer Shot' 1 and 2 architectural plan executed (step-based knowledge acquisition) for integrating UIMA, BPEL, and NwHIN CONNECT
   a) Implemented CONNECT software environment (C32 specs; CEM subsets formulate XML docs (part of meaningful use)
   b) Persistence Channels defined
      i) One Channel per model
      ii) Data stored as an XML Instance of the model
      iii) Fields extracted from XML to use as indices
      iv) XML Schema defined for each model
   c) Stored using database transactions
   d) Reviewed sources used for normalization opportunities

i) In HL7 OBR Segments
   (1) Standardize Service ID (Codes)
ii) In HL7 OBX Segments
   (1) Standardize Units
   (2) Standardize Reference Ranges
   (3) Standardize Normal Flags

e) Identified appropriate terminology services and content for infrastructure to load into our Cloud environment
   i) problems/diagnoses/signs/symptoms: SNOMED CT
   ii) laboratory observation names/identifiers: LOINC
   iii) coded values of labs: SNOMED CT
   iv) medications/drugs: RxNorm
   v) medication/drug classes: NDF-RT
   vi) procedures: SNOMED CT

DN Next Steps:
1) Continue development on the CEM browser website and launch.
2) Continue Pan-SHARP discussion, develop models, terminology, and transform.
3) Continue to align the efforts between CDISC SHARE and SHARP CEMs.

Clinical Natural Language Processing (cNLP)

The clinical and research medical community creates, manages and uses a wide variety of semi-structured and unstructured textual documents. To perform research, to improve standards of care and to evaluate treatment outcomes easily - and ideally, in an automated fashion - access to the content of these documents is required. The knowledge contained in unstructured textual documents (e.g., pathology reports, clinical notes), is critical to achieving all of these goals. For instance, clinical research usually requires the identification of cohorts that follow precisely defined patient- and disease-related inclusion and exclusion parameters.

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. Therefore, cNLP is a critical component in SHARPn. It facilitates the use of clinical narratives in the similar way as structured data for high-throughput phenotyping, decision support at the point of care, and evaluation of health care delivery outcomes to name just a few of all possible applications. In popular culture, two cNLP-based applications (albeit heavily proprietary) are the Google search engine and IBM's Watson question-answering system.

Much of the work in SHARPn surrounds cNLP methods that find the highly relevant "information nuggets" in the free text. This is usually achieved through a series of low level text processing tasks such as conducting sentence boundary detection, finding "tokens",

assigning part-of-speech tags to the tokens, finding the phrasal chunks (i.e. units of noun or preposition phrases), performing deep sentence parsing, finding the named entities and events, followed by finding the relations between events including temporality. Once clinical events and their attributes (e.g. whether the event is negated or not) are identified, SHARPn can deliver ontology maps and templates related to identified clinical events which leads to normalization. In parallel, SHARPn hones its cNLP open-source technology systems (Clinical Text Analysis and Knowledge Extraction System (cTAKES), http://sourceforge.net/projects/ohnlp/) to apply the ontology processing to any clinical note set.

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

Progress:
The SHARPn software for cNLP and IE is cTAKES available on SourceForge under a permissive Apache v2.0 license (http://sourceforge.net/projects/ohnlp/). Two main sets of 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 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 imaginable 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) Implementation of an alternative string lookup algorithm for named entity normalization
2) Concept mention detection – investigating the outcome of pooling corpora from multiple institutions
3) A coreference module combining machine learning and rule-based approaches
4) Medication annotator refinement through corpus analysis
5) Relation extraction development to identify relevant relations between clinical events
6) Annotation schema development based on CEM
7) Annotation guidelines development based on (6) and pilot annotations
8) Gold standard annotations and active learning environment for efficiently choosing the instances for gold standard annotation
9) Stratified corpus methodology
    a) Seed corpus generation
    b) Sampling methodology applied to two Electronic Medical Record systems (EMR) – Mayo Clinic and Seattle Group Health – to create a sample of 2000 clinical reports for algorithm and system development
10) A library of de-identification tools for the clinical narrative was released. This was a joint collaboration between SHARPn - MIT/SUNY team and the MITRE Corporation.

The release contains a library of integrated de-identification systems with surrogate generation.

11) Fully de-identified corpus from two contributing sites (Mayo Clinic and Seattle Group Health). Size of the corpus is 500K tokens. The corpus is being used for gold standard and algorithm development to evaluate the cNLP techniques. The corpus was created using the methodology from (9) and de-identified using the library of de-identification tools described in (10)
12) The dissemination of a cNLP algorithm requires performance benchmarking. The evaluation workbench allows cNLP investigators and developers to compare and evaluate various cNLP algorithms. cNLP developers are the targeted users for the Evaluation Workbench. In October 2011, the first iteration of the cNLP evaluation workbench was released: (http://orbit.nlm.nih.gov/resource/clinical-nlp-evaluation-workbench). The workbench GUI consists of three windows: Document pane (left), Class/Statistics pane (upper right), and Reports, Attributes and Relations panes (lower right). Nearly all behaviors of the tool are activated by moving the mouse around while holding down the control key.

Towards the goal of engineering a robust, highly scalable, user friendly, open source software, the SHARPn cNLP team had several open source releases of its main software, cTAKES. In addition, cTAKES has been adopted for the Pan-SHARP program project as the NLP platform to extract medication information from the clinical free text. As of January 5, 2012, cTAKES has over 2800 downloads. cTAKES releases are:

1) March 2011: release 1.1 included a Smoking Status module to process clinical documents and identify patients' smoking status at the patient level as well as the document level. One of five smoking status categories is associated with the patient: past smoker, current smoker, smoker, non-smoker and unknown (described in a publication; Sohn and Savova, 2009[1])
2) October 2011: release 1.2 included a SideEffect module (described in a publication, Sohn et al, 2011), which extracts physician-asserted drug side effects from clinical notes. This release also introduces an integrated version of cTAKES, icTAKES, which provides an integrated version of cTAKES for end users and developers.
3) December 2011: cTAKES 1.3 release
   a) Inclusion of a set of UMLS dictionaries (SNOMED-CT and RxNorm) in the distribution. Users no longer need to take an extra step to get production level dictionaries except to supply a UMLS username and password.
   b) Two new modules - Constituency Parser and Coreference resolver (described in a publication, Zheng et al, In press)

Because cTAKES is software developed collaboratively across sites following the Apache Software Foundation philosophy, several engineering decisions have been made to enable the eco-system. The first engineering decision is related to the cTAKES data model referred to as the Type System. The type system is used as a starting point for any subsequent application be it Phenotyping, patient care, comparative effectiveness, meaningful use of

[1] Sohn S, Savova G. 2009. Mayo Clinic smoking status classification system. Proc. AMIA, Nov. 2009.

the EMR, etc. Two other enabling engineering decisions are code governance structure and code sharing environment. The former is facilitated by the cNLP Software development group, while the latter is provided by the SHARPn cloud.

Notable Milestones Reached in cNLP:

1) Patient Smoking status module in cTAKES 1.1 (March 2011)
2) Clinical Element Model OrderMedPopulation module (May 2011)
3) Full-cycle pipeline (free text-NLP-CEM) as part of the SHARPn v1 pipeline (June 2011)
   a) All medications in Mayo Clinic data set extracted with cTAKES (NLP method)
   b) Processed 360,452 notes for 10,000 patients
   c) 3,442,000 CEMs were created; Processing time: 1.6 sec/doc
4) Side effect module in cTAKES 1.2 (October 2011)
5) Release of cTAKES with an integrated cTAKES option (icTAKES) (October 2011)
6) Evaluation workbench, v1 (October 2011) http://orbit.nlm.nih.gov/resource/clinical-nlp-evaluation-workbench
7) Constituency parser module in cTAKES 1.3 (December 2011)
8) Coreference module in cTAKES 1.3 (December 2011)
9) cTAKES 1.3 with an integrated UMLS database (December 2011)
10) A set of tools for tailoring the UMLS through corpus analysis
11) Aligning annotations with ISO Linguistic standards and other conventions
12) Identification of a set of eventive UMLS entity types and relations for the automatic extraction of medical events
13) Enhanced cTAKES type system
14) MIT/SUNY v1 of de-identifier and surrogate generator
15) 150K token seed corpus generation and de-identification
16) Annotation model and schema based on Clinical Element Models for diseases/disorders, signs/symptoms, medications, anatomical sites, procedures, labs
17) Annotation guidelines finalized

cNLP Next Steps:

1) March/April 2012: a major release of cTAKES with new components for normalization against the Clinical Element Model templates. This will constitute version 1 of the overall final deliverable from the NLP team.
2) Application of the side effect module for health care quality research.
3) Application of cTAKES for the Pan-SHARP project for Medication Reconciliation.
4) Hybrid engineering approach for packaging cTAKES – separation between core NLP modules and application modules – including the implementation of the Enhanced Type System.
5) Gold standard annotation of the first 150K tokens.
6) Full testing of a fast string lookup algorithm for entity mention discovery (using (5)).
7) Improved module to discover negated and hedged clinical events.
8) Module to discover the experiencer of a clinical event.
9) Refinement of the coreference module combining machine learning and rule-based approaches (using (5)).
10) Semantic role labeler trained on clinical data (using (5)).
11) Clinical Events Relation extraction module trained on clinical data (using (5)).

12) Improved Medication extraction module (using (5)).
13) A library of Medication extraction modules.
14) Integration of a tokenizer updates from Multi-source Integrated Platform for Answering Clinical Questions (MiPACQ) project to cTAKES to upgrade to the latest Penn Treebank guidelines. Retraining of downstream components. This integration will allow linking cTAKES to general domain search engines to enable even deeper IE through mining the literature, patient blogs, and the web.
15) Integration of the Term Expander from MiPACQ project to cTAKES. This will enable better synonymy matching.
16) Enhanced Evaluation Workbench.

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:
In the area of phenotype processes, many projects such as SHARPn, Researchers in the Electronic Medical Records and Genomics (eMERGE) network, Informatics for Integrating Biology and the Bedside (i2b2) and other projects are doing “phenotyping” (broadly speaking). This is resulting in the creation of a large number of robust, validated algorithms in which most are represented in MS Word files, JPEG pictures, etc. and currently there is no standard format or designated “home” for these algorithms. The SHARPn team has been working on a “PhenoPortal” or electronic, publically accessible portal and library for such algorithms with core requirements identified. The algorithm criterion will be published in a standards-based representation, anchored to CEMs where possible and be machine executable for users.

It is the “PhenoPortal” work that has brought about collaboration with the National Quality Forum (NQF) and investigation of the Measure Authoring Toolkit (MAT). The MAT is a

Web-based application that allows you to create eMeasures without writing XML. The SHARPn team plans to author the Diabetes and Hypothyroidism criteria using Quality Data Model (QDM). Phenotype Team submitted request to get Access of source code as well as user authority MAT Quality forum on November 22, 2011 and finally got approval on December 5, 2011. The team worked on MAT tools to create phenotype eMeasure but found that the tool is not ready yet and less user friendly. Problem had been submitted on December 14, 2011 with some sample codes and got response from Ahmed on December 21, 2011, to have a meeting in January 2012.

Working with disease specific algorithms, the SHARPn team has focused on identification and modification of existing, and where applicable, creation of new CEMs for EHR based phenotyping algorithms (Peripheral Arterial Disease (PAD), Type 2 Diabetes (T2D), Hypothyroidism, and Community Acquired Pneumonia). Analysis of "gap" between existing CEMs, and what is required for the phenotyping algorithms was conducted and fed into the core CEM work noted under the data normalization activities.

In collaboration with the data normalization team, the HTP team led a project to focus on a formal definition of the CEM using semantic web specifics. They identified representative models from the CEM repository at Intermountain Healthcare and conducted a feasibility of RDF representation of CEMs. A manuscript of RDF-based representation of CEM was accepted for publication at the 2011 AMIA Fall Symposium, Washington DC.

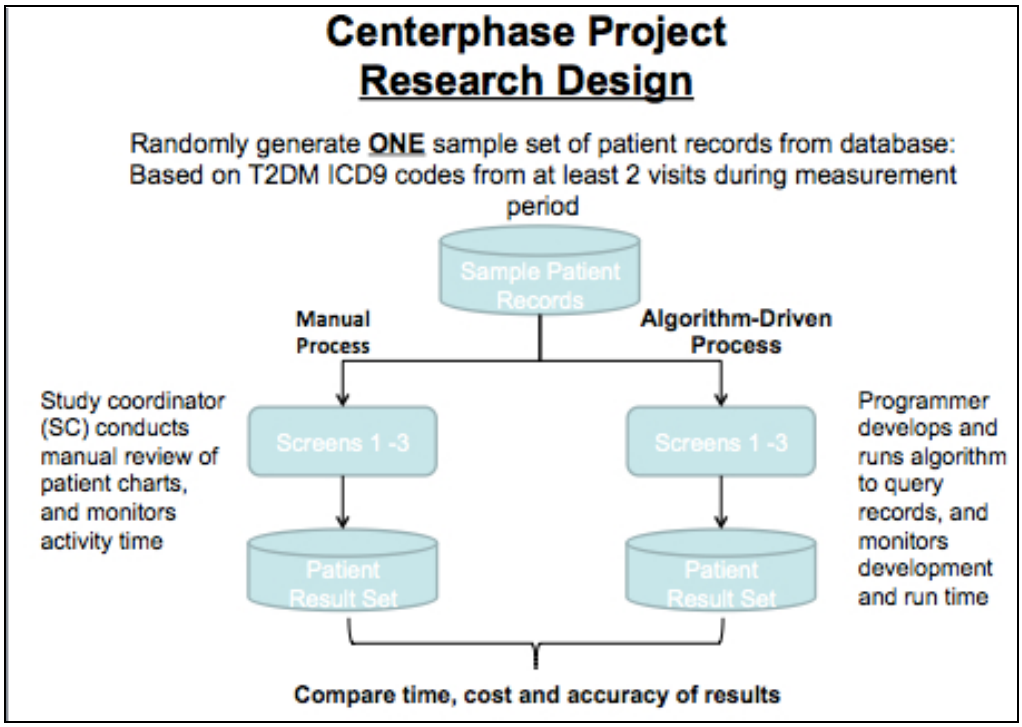
In addition to RDF, the HTP team has focused on leveraging a standards-based protocol representation model for structured modeling of phenotyping algorithms. An analysis of Analysis of eMERGE network phenotyping algorithms for identifying commonalities, differences, data elements used, phenotyping logic used, etc. was conducted and their representation in CDISC protocol representation model, Drools, GELLO (and NQF) were conducted on three of the algorithms. Evaluation results available for 13 eMERGE algorithms with respect to data elements used, terminologies used, and phenotyping logic for representation of the algorithms was accepted as a 2011 AMIA Fall Symposium Manuscript.

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.

Extending beyond the JBoss application server, the team is exploring a more visual interface for its algorithm execution. PopHealth (http://projectpophealth.org) is a likely model and will be explored in 2012. PopHealth is an open source reference implementation software service that automates the reporting of Meaningful Use quality measures. popHealth integrates with a health care provider's EHR system using continuity of care records. It has released new features such as practice-level reporting with multiple providers, report stratification on race/ethnicity/spoken language/gender, individual

patient view, with meaningful use clinical gap analysis from Continuity of Care Document (CCD) data, clinician-gated, manual override support for exclusion logic.

To demonstrate the market value of EHR-derived phenotyping, the Centerphase team is studying the financial aspects of manual phenotyping vs. electronic phenotyping with the medicinally managed diabetes algorithm using the proposed SHARPn tools and technologies. Early in 2011 the team designed an approach to determine market metrics for EHR derived phenotyping algorithms, developed the use cases and facilitated a pilot study of 50 subjects. The team is currently in the process of a larger co-hort evaluation and draft manuscript.



Notable Milestones Reached in HTP:

1) Prototype implementation of eMERGE Diabetes and Hypothyroidism algorithm using Drools
2) Access to the NQF Measure Authoring Toolkit and authoring of Diabetes and Hypothyroidism criteria using QDM
3) Set up JBoss and Drools environment within the SHARPn cloud infrastructure
4) Prototype software for translating NQF based criteria into Drools executable rules
5) Local installation of popHealth open-source framework
6) Extended collaborations with eMERGE, PGRN and CTSA colleagues and assist in the design and development of phenotyping algorithms

HTP Next Steps:

1) First release of NQF to Drools convertor (open-source UIMA-based annotator)

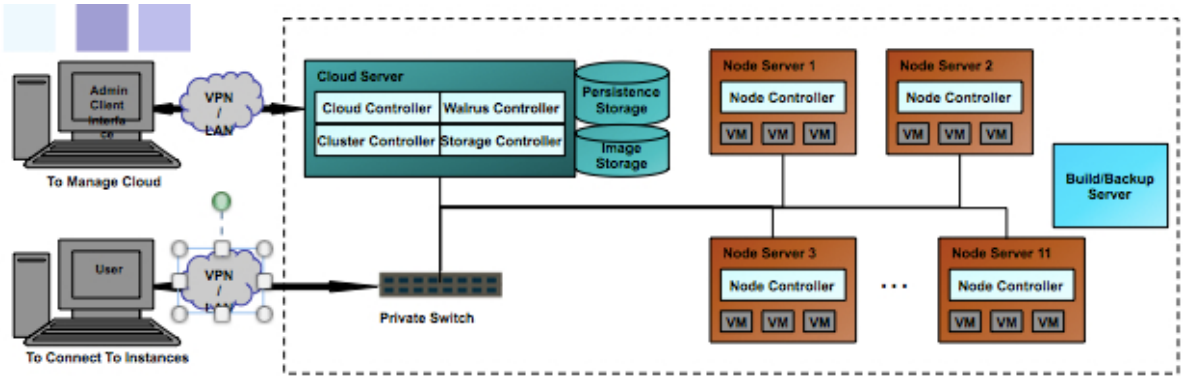
2) Investigations to modify/adapt popHealth infrastructure for developing the graphical interface for executing phenotyping algorithms
3) Early prototyping efforts for design and implementation of PhenoPortal – a portal for authoring, visualization and execution of phenotyping algorithms

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. This project ensures that these software resources can scale to near-real-time performance or operate as standalone applications. In SHARP Area 4 the IBM Watson Unstructured Information Management Architecture (UIMA) technology framework is the backbone of the various tools and architecture.

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:
The SHARPn cloud computing environment, also known as "the cloud" is proudly constructed by the SHARPn infrastructure team and is basically a set of virtual machine images that can be instantiated, used, and shut down for Secondary Use of EHR Data research. The file share server has about 7TB of space currently set up. The software developed by SHARPn will allow researchers and clinicians to pull together very different types of health care data and ask questions about disease, prevention, outcomes, and care delivery. Answering these questions requires a robust, high-quality data set which can be generated by software in the SHARPn cloud computing environment.



Within the cloud computing environment there are several primary goals:

- Sharing data with the consortium

- Clinical Data Normalization Services and Pipelines
- Natural Language Processing
- High-throughput Phenotyping

And many virtues:

- Secure infrastructure
- Scaling Capacity
- Virtualized services (reusable and agile)
- Highly available, common infrastructure

Currently the Cloud is hosting 10 images., examples of the most recent activity includes:

- emi-DD121096 - LexEVS 6.0, MySQL 5.1 64 bit, Java 1.6, OS - Ubuntu 11.04
- emi-F08E10CA - MySQl 5.1 64 bit, OS - Ubuntu 11.04
- emi-3FEA1272 - NwHIN Aurion 4.0, Mirth Connect 2.1.0.5374, GlassFish 2.1, NetBeans IDE 6.7.1 SoapUI, Metro 1.5, MySQL 5.1.55 64 bit, Java JDK 1.6.0_21, OS - Ubuntu 11.04
- emi-277E121D - cTakes 1.2.1, Mirth Connect 2.1.0, Java JDK 1.6.0_21, OS - Ubuntu 11.04
- emi-406D12E0 - Ubuntu 11.04 64 bit Desktop
- emi-D0E6153C - JBoss 6.0, Drools 5.2.0, Eclipse Hello IDE 3.6.2, Maven 3.0.2, Java JDK 1.6.0_21, OS - Ubuntu 11.04

With the release of the cloud, SHARPn has generated end-user documentation on the base cloud infrastructure with 11-wiki pages of content.

The SHARPn work has worked to make the best use of the UIMA framework and has worked directly with IBM on the application requirements for scaleout, deployment, and parallelization, and help with the scaled-out platform design. IBM Research's part, during this period, is the beginning of a new set of UIMA launching & monitoring framework capabilities - to help manage large scale (e.g., cloud-based) infrastructure running UIMA jobs. On the UIMA side, during this period there were two Apache UIMA releases: 2.3.1 of the "Add-ons" package, and 2.4.0 of the base UIMA Java SDK; of which the SHARPn IBM resource was the release manager: http://uima.apache.org/news.html#29 August 2011 and http://uima.apache.org/news.html#07 December 2011. IBM also deployed two consultants to Mayo Clinic to do an architectural review of the SHARPn UIMA design and provide expert opinion on next step improvements.

In establishing the pipeline, SHARPn collaborated closely with Mirth. This effort was led by Agilex. They conducted the following critical tasks:

1) Developed, deployed and tested the Mirth Channel for sending HL7 lab messages using the XDR Connector. This channel leverages the NwHIN XDR adaptor capabilities.
2) Developed, deployed and tested the Mirth Channel for receiving HL7 lab messages from the NwHIN using the XDR capabilities.
3) Tested NwHIN Aurion XDR Adaptor (development was done during 1st half of 2011).
4) Tested UIMA connector capability within Mirth Connect with assistance from Mirth Team (development was done during 1st half of 2011).

5) Created deployment files and installation documentation for the Tracer Shot 1 and Tracer Shot 2 demonstrations. Worked closely with Mayo Clinic and IHC in installing and configuration needs prior to each demonstration and test.
6) Developed, tested, and deployed CEM Administrative Diagnosis to DB Channel – receives an XML instance of an Administrative Diagnosis CEM and persists that to the CEM Relational Database. (Some of this development was done during 1st half of 2011).
7) Developed, tested, and deployed CEM Lab to DB Channel – receives an XML instance of a Lab CEM and persists that to the CEM Relational Database. (Some of this development was done during 1st half of 2011).
8) Developed, tested, and deployed CEM Medication to DB Channel – receives an XML instance of a Medication CEM and persists that to the CEM Relational Database. (Some of this development was done during 1st half of 2011).

As proof-of-concept in the SHARPn pipeline process, two "Tracer Shot" pilots have been successfully completed. In Tracer Shot 1, conducted in mid-2011: The team worked closely with a number of the SHARP members on completing the phase 1 demonstration for taking HL7 2.x messages (Lab, Medications, and Administrative Diagnosis data) and sending that data over the NwHIN network from Intermountain Healthcare to the Mayo cloud, having that data normalized via the UIMA pipeline and then storing the data into a normalized data store using the CEM data models and Mirth Connect. An NwHIN Gateway instance was set up on the cloud to receive and send messages between Intermountain Healthcare and Mayo Clinic. A cTAKES instance was set up on the cloud to process incoming Lab, Diagnosis and Medication data from Intermountain Healthcare and Mayo. A simple relational database based on the Schema design and CEM requirements from Intermountain Healthcare. This data store was used to store the Lab, Medication, Demographic and Administrative Diagnosis information after it had been normalized via the UIMA Data Normalization pipeline. Agilex created several Mirth Channels that would receive the CEM data from the UIMA pipeline and then store that data into the CEM Relational Data store. The pilot resulted in approximately 2.5 million patient records stored. A JBoss Drools instance was set up on the cloud to access and run the phenotype algorithm rules for medicinally managed diabetes against this CEM database.

Leveraging the lessons learned from the first phase, a Tracer Shot 2 architecture enhancement began late in 2011 focused on Lab data. The Agilex team worked closely with a number of the SHARP members on completing this phase 2 demonstration for taking HL7 2.x messages (Lab Only) and sending that data over the NwHIN network from Intermountain Healthcare to the Mayo cloud without any middle-man manual movement of the data. They installed and deployed updated Mirth Channels with the enhancements determined during the Tracer Shot 1 demonstration. They created a set of test data and testing scenarios using de-identified data from Intermountain Healthcare and facilitated the successful testing of these between Intermountain Healthcare and Mayo Clinic.

The corresponding role of the Infrastructure team is to tackle issues that arise. Below represents the issues encountered and their resolution status from the support team.

1) Server room AC outage:

a) In mid-September, the temperature inside the server room went very high causing one of the power channels to trip. This power channel trip caused many of our node servers to shutdown, disrupting many running instances. This accidental shutdown made the cloud unusable. Just restarting the cloud did not help many volumes were not able to be mounted.
b) To fix the cloud, we had to upgrade all the servers from Ubuntu 10.04 to Ubuntu 11.04. Mayo Clinic Facilities fixed the AC problem in the server room.
c) Steps have been taken regarding this issue: All our servers were distributed across four power circuits in the server room.
d) Staff created a communication process document in order to monitor/report air temperature. A wireless thermometer has been placed inside the server room; the thermometer allows monitoring of the temperature inside to ensure the room does not get too warm and cause damage to the servers. Staff on the floor contact us immediately when temps near 75 degrees. Any power or chilled water outage which might affect the cooling system is monitored.

2) Power Outage:
   a) Unexpected power outage occurred, which lasted for a second or two causing our front facing server, cirrusmatic.mayo.edu, to shutdown resulting in disrupting all the running instances.
   b) Two new UPS systems have been ordered. A maintenance outage to bring down the cloud to connect all the servers to UPS systems is scheduled. This might prevent disrupting our cloud in case of unexpected brief power outages.
3) Memory leak in Eucalyptus 2.0:
   a) There is a known memory leak issue with Eucalyptus (software that provides cloud functionality) 2.0. This has occurred three times. One of the Eucalyptus services was using up all the memory and not releasing it. Once all the memory is used, their HyperSQL database which runs in the memory shuts down. This causes all administrative functions like starting an instance, creating new users, attaching a volume, etc. to be unavailable. All three times, this issue occurred 2 weeks from the cloud restart.
   b) Communication with Canonical to come up with a solution to this issue has happened. They recommended some configuration changes which we applied. However, this may not fix the problem 100%. The Eucalyptus team has the fix for this issue in their upcoming release 3.0 which will happen around April 2012. Meanwhile, a monitoring system has been installed on our cirrusmatic server to monitor and record resource usage statistics.
4) VPN connection:
   a) This is not directly related to the cloud infrastructure, but affected access to the cloud by a few users from IHC and Harvard.
      i) IHC: User had problem accessing the cirrusdata file server due to firewall at IHC. Worked with that user and the IHC IT team to resolve this issue.
      ii) Harvard: Some of the users in Harvard use Ubuntu Desktop machines and were having problem connecting to Mayo Clinic's network using Mayo Clinic's web based F5 VPN Network. We checked with Mayo Clinic's VPN team and learned that Ubuntu is not supported. Another approach to access Mayo Clinic Network

is using F5 client. This information has been forwarded to the Harvard team to try this approach.

Next Steps:

1) The team is currently working on next iterations of Tracer Shot development and Pan-SHARP execution, which will include XDR channels to support Medication and Administrative diagnosis messages with the inputs and outputs being configurable directories that Intermountain Healthcare and Mayo Clinic can use for the next set of testing scenarios.
   a) XDR Channels for medication and diagnosis messages
   b) Single CEM for multiple OBX segments
   c) Efficiently utilize terminology services (integrate code maps, e.g. LOINC)
   d) Incorporate a library for HL7 clean-up routines

Data Quality (DQ)

Having normalized data does not help if phenotyping encounters conflicting or inconsistent data. This project provides high-confidence algorithms and services to detect and optionally reconcile such data.

DQ Aims: Develop statistical profiles of: a) malformed data (failing transformation checks), b) non-semantic data (failing vocabulary profiles), c) inconsistent data (failing phenotype specific profiles), and d) conflicting data (lab or medicine characteristics incompatible with diseases, and the presence of negation and assertion for the same elements). These profiles will include frequencies, proportions, and variance measures. Create statistically based confidence measures that will be reported to the UIMA pipeline, enabling users to dynamically parameterize thresholds for rejection of spurious data.

Progress:
In a heterogeneity study, the team is researching the variation across health care delivery organizations in the way medical information is recorded and coded. In order to develop valid methods for combining information across health care delivery organizations, it is important to understand the variation in the EHR data relative to any given disease or phenotype. 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.

IRB approvals to extract and share de-identified data between Intermountain Healthcare and Mayo Clinic collaborators have been approved at both institutions. SQL code has been prepared for data extraction with each organization currently pulling preliminary data; all patient records that are usable for research purposes would be accessed for adult (18+

years old) subjects who are represented in the database during the period from 2009-2010. No Natural Language Processing would be attempted in this pilot study. The resulting data would be placed into a combined database, with a patient study number replacing the institutional patient ID number.

In phase 1, we will use machine learning association methods to discover the codes, based on an initial list of codes generated by the Northwestern eMERGE T2DM algorithm, that associate together with each other and with the phenotype of interest. This approach removes bias due to preconceived expert opinions on what might or might not be important to retrieve. This phase may be conducted on a sample.

In phase 2, every code or procedure from these lists would be extracted along with date and result and saved in a non-identifiable manner. The resulting database would be analyzed by comparing frequencies of each code or finding across institution, both over the entire respective populations, but also within a subpopulation defined as having some evidence of T2DM from any of these findings. These results would be used to infer whether coding practices, utilization of tests, and recording of results are comparable across the two organizations.

Data Heterogeneity Study Next Steps:

1. Refinement of screening algorithm at both institutions
2. Analysis
    - Univariate Comparison of frequencies (code by code, item by item)
    - Comparison of relative frequencies (between codes, between items, between findings)
    - Sensitivity to time interval (compare 2-year and 1-year results)
    - Associations between demographics and frequency of items
    - Associations between health care access characteristics and frequency of items
3. Manuscript publication/white paper
    - Analysis of frequency comparisons - manuscript (Dr. Kent Bailey)
    - Analysis of associations among data - manuscript (Susan Welch)

In a second study, the team is focused on Body Mass Index (BMI) to determine availability, location, and accuracy of data on height, weight, and BMI, in the EHR, and impact of this on phenotyping algorithms. The aim is to apply knowledge discovery methods to suspect errors to identify correlates of errors. IRB approvals for preliminary BMI data quality study data extraction has been obtained by Intermountain Healthcare and Mayo Clinic collaborators. A literature review for EHR BMI data is underway.

BMI Study Next Steps:

1. Developed SQL code (using heterogeneity study data above) to:
    - Extract height, weight and BMI data and correlates per literature, demographics and health status
    - Generate aggregate descriptive statistics over study cohort.

- Identify observations as 'errors' according to published algorithm[2]
- Correlate algorithm errors with other measures of variation
- Associations of BMI/obesity markers with item frequencies

2. Analysis
3. Apply knowledge discovery methods to suspect errors to identify correlates of errors
4. Manuscript publication/white paper
   - Identification of suspect BMI data: Implications for Shared Secondary Use
   - Correlates of BMI Errors in the EHR

Evaluation Framework

Evaluation of the SHARPn pipeline system for securely collecting and normalizing clinical data for secondary use is based on two approaches. The first approach involves the end-to-end testing of the system to demonstrate that data can be transferred from one site to another, can be translated into a standard structure, can be translated into a standard terminology, can be validated against a datatype-specific model, can be normalized to use standard datatypes and units of measure, and can be stored in a database for use in research, quality assurance, and other relevant activities. The second approach will involve demonstrating successful use of this data as in accurate identification of patient phenotypes using data from multiple institutions.

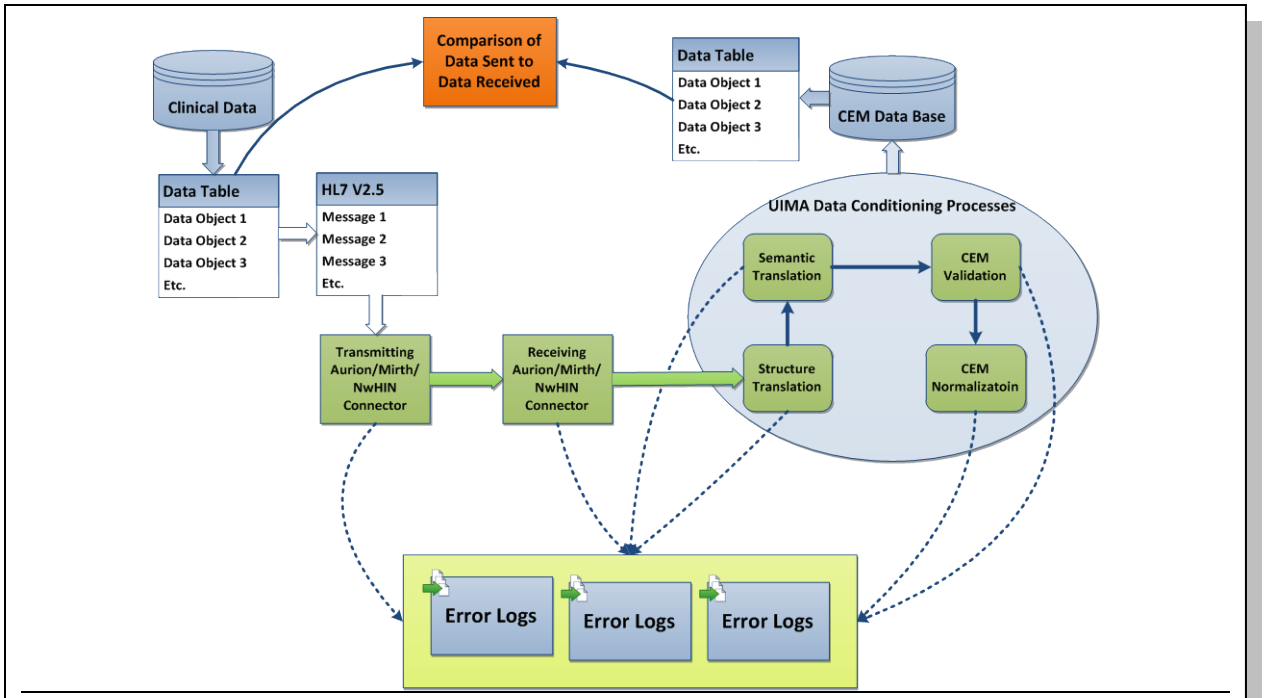


Figure 1: Conceptual diagram of the SHARPn data transfer and translation infrastructure. Testing of this system will involve transfer of a collection of data from one side to another, processing of that data to give it a standard structure and semantics, storage of the data in a database, and then extraction of a data set to compare with the originally transmitted data.

[2] Goldberg SI, Shubina M, Niemierko A, Turchin A. A Weighty Problem: Identification, Characteristics and Risk Factors for Errors in EMR Data, AMIA Annu Symp Proc 2010 (2010) 251-255.

The initial evaluation is designed as a formative process in which evaluation is done to identify deficiencies and these deficiencies are addressed in the next development cycle. Subsequently, the system is tested again to verify the success of the enhancements. Each new cycle of testing and development results in further improvements to these tools. Figure 1 illustrates the testing process that has been designed and which we are implementing. The goal is to demonstrate that after transmission, translation, validation, and normalization the information received still corresponds to the information sent.

The second type of evaluation tests the outcomes of the phenotyping process. Phenotypes are generated using the initial, pre-transmission data set and then the same phenotyping rules are used to process the final data set after transmission, translation, validation, and normalization. Discrepancies in the two results demonstrate the effect of failures in the transmission pipeline.

Milestones Reached:
1) Key infrastructure is available to begin the evaluation process.
2) Initial data sets have been extracted from the Intermountain Healthcare Enterprise Data Warehouse (EDW), have been de-identified, and have been configure as streams of HL7 messages.
3) Transmission of messages has been demonstrated.

Next Steps:
1) The CEM database is being reconfigured to better support storage of the normalized data.
2) A comprehensive approach to maintaining and accessing the component error logs is being developed. Once the steps are complete, the initial test of the system can occur.

Program Outputs

a) Products

i) Data Normalization / CEM Database Design
   (1) Models and terminology for demographics, labs, drugs, and disorders
   (2) CEM request website

ii) NLP
   (1) UIMA Type system for NLP (enhanced cTAKES Type System)
   (2) MIT/SUNY de-identification tool as part of the SHARPn library
   (3) Draft document summarizing lessons and recommendations from the cloud security roundtable
   (4) Annotated data
   (5) Evaluation workbench, v.1
   (6) CEM OrderMedPopulation, May 2011
   (7) March 2011: cTAKES release 1.1 included a Smoking Status module to process clinical documents and identify patients' smoking status at the patient level as well as the document level. One of five smoking status categories is associated

with the patient: past smoker, current smoker, smoker, non-smoker and unknown (described in a publication; Sohn and Savova, 2009[3])

(8) October 2011: cTAKES release 1.2 included a SideEffect module (described in a publication, Sohn et al, 2011), which extracts physician-asserted drug side effects from clinical notes. This release also introduces an integrated version of cTAKES, icTAKES, which provides an integrated version of cTAKES for end users and developers.

(9) December 2011: cTAKES 1.3 release

(a) Inclusion of a set of UMLS dictionaries (SNOMED-CT and RxNorm) in the distribution. Users no longer need to take an extra step to get production level dictionaries except to supply a UMLS username and password.

(b) Two new modules - Constituency Parser and Coreference resolver (described in a publication, Zheng et al, In press)

iii) Infrastructure

(1) End-to-end pipeline, v1, June 2011

(2) End-to-end pipeline, v2, December 2011

(3) Mirth Channels

(a) Sample XDR Channel – to push data via NwHIN Gateway

(b) ReceiveXDRMessage – to receive data via NwHIN Gateway

(c) CemAdminDxtoDatabase – Store Billing Codes to CEM Database

(d) CemLabToDatabase – Store Labs Results to CEM Database

(e) CemMedicationToDatabase – Store Meds to CEM Database

b) Publications and Presentations

i) Recent/accepted/published

(1) Aberdeen J. NLP techniques for clinical record de-identification, presentation to AcademyHealth Annual Research Meeting, Seattle, June 12-14, 2011.

(2) Chapman W, Nadkarni P, Hirschman L, D'Avolio L, Savova G, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of American Medical Informatics Association. 2011 -:1e4. doi:10.1136/amiajnl-2011-000465.

(3) Choi J, Palmer M. Getting the most out of Transition-based Dependency Parsing, In the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL-HLT 2011, June 19 - 24, 2011, Portland, OR.

(4) Choi J, Palmer M. Transition-based Semantic Role Labeling Using Predicate Argument Clustering, In the Proceedings of RELMS 2011: Relational Models of Semantics, held in conjunction with ACL-HLT 2011, June, 2011, Portland, OR.

(5) Chute CG, Pathak J, Savova GK, Bailey KR, Schor MI, Hart LA, Beebe CE, Huff SM. The SHARPn Project on Secondary Use of Electronic Medical Record Data: Progress, Plans and Possibilities. AMIA 2011 (paper).

(6) Clark C. Recent efforts in clinical NLP: Uncertainty discovery through NLP, presentation to Natural Language Processing Workshop, i2b2 Academic Users Group, Boston, June 28, 2011.

[3] Sohn S, Savova G. 2009. Mayo Clinic smoking status classification system. Proc. AMIA, Nov. 2009

(7) Conway MA, Berg RL, Carrell D, Denny JC, Kho AN, Kullo IJ, Linneman JG, Pacheco JA, Pessig PL, Rasmussen L, Weston N, Chute CG, Pathak J. Analyzing Heterogeneity and Complexity of Electronic Health Record Oriented Phenotyping Algorithms. AMIA 2011 (paper).
(8) Conway MA, Pathak J. Analyzing the Prevalence of Hedges in Electronic Health Record Oriented Phenotyping Algorithms. AMIA 2011 (poster).
(9) Dligach D, Palmer M. Good Seed Makes a Good Crop: Accelerating Active Learning Using Language Modeling. In the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL-HLT 2011, June 19 - 24, 2011, Portland, OR.
(10) Dligach D, Palmer M. Reducing the Need for Double Annotation. In the Proceedings of the Fifth Linguistic Annotation Workshop (LAW V) held in conjunction with ACL-HLT 2011, June, 2011, Portland, OR.
(11) Hirschman L. Evaluation as a driver in Software Communities, presentation to Workshop on Designing an Ecosystem for Clinical NLP, Integrating Data for Analysis, Anonymization and Sharing (iDASH), University of California, San Diego, May 2-3, 2011.
(12) Liu H, Wagholikar K, Wu S. Using SNOMED CT to encode summary level data - a corpus analysis. AMIA CRI 2012.
(13) MITRE System for Clinical Assertion Status Classification, JAMIA 2011; Published Online First: 22 April 2011 doi:10.1136/amiajnl-2011-000164.
(14) Rea S, Pathak J, Savova GK, 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. Second stage of review at JAMIA.
(15) Savova G, Olson J, Murphy S, Cafourek V, Couch F, Goetz M, Ingle J, Suman V, Chute C, Weinshilboum R. The electronic medical record and drug response research: automated discovery of drug treatment patterns for endocrine therapy of breast cancer. Journal of American Medical Informatics Association. 2011.
(16) Savova GK, Chapman WW, Elhadad N, Palmer M. 2011. Shared annotated resources for the clinical domain. AMIA ann symp. Panel.
(17) Sohn S, Kocher J-P, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. JAMIA 2011; 18:i144-i149.
(18) Sohn S, Wu S. Dependency Parser-based Negation Detection in Clinical Narratives. AMIA CRI 2012.
(19) Tao C, Parker CG, Oniki TA, Pathak J, Huff SM, Chute CG. An OWL Meta-Ontology for Representing the Clinical Element Model. AMIA 2011 (paper).
(20) Tao C, Welch SR, Wei WQ, Oniki TA, Parker CA, Pathak J, Huff SM, Chute CG. Normalized Representation of Data Elements for Phenotype Cohort Identification in Electronic Health Record. AMIA 2011 (poster).
(21) Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. JAMIA 2011 Sep-Oct; 18(5) 580-7

(22) Wagholikar K, Torii M, Jonnalagadda S, Liu H. Feasibility of pooling annotated corpora for clinical concept extraction. AMIA CRI 2011
(23) Wu ST, Kaggal VC, Savova GK, Liu H, Dligach D, Zheng J, Chapman WW, Chute CG. Generality and Reuse in a Common Type System for Clinical Natural Language Processing Proceedings of the First International Workshop on Managing Interoperability and compleXity in Health Systems. Glasgow, Scotland. 2011.
(24) Wu S, Liu H. Semantic Characteristics of NLP-extracted Concepts in Clinical Notes vs. Biomedical Literature Proceedings of the Annual AMIA Fall Symposium. Washington DC. 2011.
(25) Wu S, Liu H, Li D, Tao C, Musen M, Chute CG, Shah N. UMLS Term Occurrences in Clinical Notes: A Large-scale Corpus Analysis. AMIA CRI 2012.
(26) Wu S, Wagholikar K, Sohn S, Kaggal V, Liu H. Empirical Ontologies for Cohort Identification. Text REtrieval Conference. 2011.
(27) Zheng J, Chapman W, Miller T, Lin C, Crowley R, Savova G. In Press. A system for coreference resolution for the clinical narrative. Journal of the American Medical Informatics Association.

i) Planned
(1) Manuscript/s on the comparison of different dictionary lookup approaches
(2) Manuscript/s on active learning methodology
(3) Manuscript on gold standard annotations
(4) Manuscript, "When NLP tools meet the medical world" describing porting parsing tools from the general to the clinical domain
(5) Manuscript/s on comparison of coding, lab, and med data between institutions
(6) Manuscript/s on cost-benefit analysis of computer-assisted screening of EHR for phenotype identification or high-risk status identification
(7) Manuscript/s specifying mapping of CEMs to Use Case algorithms' input specifications: focusing AMIA Clinical Trials or Translational Informatics conference.
(8) Manuscript for submission to JBI publication focusing application and use of standards with input/authorship from among the SHARPn team.

Events

12/13/2011-12/15/2011: All ONC Grantee Meeting, Washington DC
Dr. Christopher Chute and Lacey Hart represented SHARPn at the All ONC Grantee meeting. Dr. Christopher Chute presented at a breakout session.

1/4/2011: Basis Technology; Cambridge, MA
Prof. Peter Szolovits gave a talk at Basis Technology, Cambridge MA.

3/21/2011: ReID Software Overview Webinar
ReID Software Overview Webinar - Details: file formats, surrogate approach taken for all PHI Types, invocation of application and support utilities. Presented by Ira Goldstein and Ken Burford.

3/29/2011: Distinguished Lecture Series; Wayne State University; Detroit, MI
Prof. Peter Szolovits lectured in the Distinguished Lecture series at Wayne State University, Detroit, on March 29.

4/5/2011-4/7/2011: ISO TC 37 standards meeting; Boston, MA
Objective: Standards for layers of linguistic annotations. Martha Palmer and Guergana Savova are members of ISO TC 37.

5/2/2011: Developing an NLP Ecosystem; UCSD, La Jolla, CA
Objective: Understand the disparate and overlapping initiatives in developing and disseminating shareable NLP tools and annotated data and determine what gaps exist between current efforts and an ideal software/data ecosystem. Our contribution: oral presentations by Christopher Chute, Guergana Savova, Wendy Chapman, Lynette Hirschman, David Carrell. Our attendees: Christopher Chute, Guergana Savova, Wendy Chapman, Peter Haug, Hongfang Liu, Ozlem Uzuner, Lynette Hirschman, Cheryl Clark, David Carrell.

5/23/2011-5/24/2011: Security Roundtable for Cloud-Deployed Clinical Natural Language Processing; Seattle, WA
This roundtable brings together nationally-recognized experts in information security, key stakeholders from health care and research institutions, and representatives of leading cloud service providers to identify legal, regulatory, technical and governance prerequisites for secure, regulatory-compliant processing of patient clinical information in externally-hosted computing environments.

6/28/2011: NLP workshop at the i2b2 Academic User Group Meeting; Boston, MA
Presentations by Cheryl Clark and Guergana Savova.

6/30/2011 – 7/1/2011: Area 4 SHARP Face to Face Conference; Rochester MN
89 Attendees with representatives from SHARPn, other SHARP programs, ONC, FSC, PAC, and Beacons.

7/2/2011: Learning from Clinical Text; International Conference; Seattle WA
Prof. Peter Szolovits gave a keynote address at the Workshop on Learning from Clinical Text at the International Conference on Machine Learning, Seattle.

7/11/2011-7/12/2011: SHARPFest; Washington, DC
Attendees: Christopher Chute, Stan Huff, Lacey Hart, Jyoti Pathak, Guergana Savova.

10/19/2011: The Quality Data Model (QDM): Version 2011
Jyoti Pathak, Dingcheng Li and Gopu Shrestha attended this Webinar to be aware about QDM and its new features. It is a free webinar on the Measure Authoring Tool Update. It has discussed how we can create and maintains multiple versions of a measure, Search and reuse value sets created by other users etc. The key Speaker was Floyd Eisenberg, Senior Vice President, Health Information Technology NQF. The purpose was to know more about QDM and its Measuring tool.

11/2011: MDHT Tools Demonstration
Les Westberg prepared and gave an overview presentation on the MDHT tools and how these tools could potentially be used by the SHARP and BEACON projects for working with CCD type document payloads for data sharing and data normalization.

11/8/2011: IBM/Mayo Clinic Visit – Review of UIMA Pipeline; Rochester, MN
IBM staff visited 41st Street Professional building to be aware the SHARPn program and review of technical approach. Jyoti Pathak, Dingcheng Li and Gopu Shrestha oriented the Technical details about the Phenotype Project to IBM Staff. It was a brief overview about the tools to be used such as QDM/MAT for Authoring tool and JBoss Drools (Guvnor) for inference and Analysis.

11/9/2011-11/11/2011: ApacheCon; Vancouver, Canada
Purpose of event was to facilitate networking with other members of the Apache open source community, and put people and faces with names we only knew via mailing lists. A secondary purpose was to investigate/discuss with others issues around OSGi enablement; UIMA has some OSGi support, and will likely have more in the future.

12/7/2011: UMLS Concepts and Terms in Clinical Notes: Large-scale Corpus Analysis; NCBO Webinar
This series aimed to showcase new projects, technologies and ideas in biomedical ontology Presented by Stephen Wu.

12/8/2011: Implementing Electronic Measures (eMeasures) for Hospitals
This webinar highlighted the importance of eMeasures in quality measurement, EHRs, and health care delivery. Discussed major issues (e.g., work flow, product evaluation) that hospital systems should contemplate when planning to implement eMeasures. Discussed the anticipated role of and relationship between hospitals and EHR vendors in implementing eMeasures. Gopu Shrestha attended this webinar and the purpose was to understand more about eMeasures role and its implementation.

12/15/2011: caGrid User Group Teleconference
This teleconference has talked about The Federated Aggregate Cohort Estimator (FACE) - CTSA Administrative Supplement its collaboration of UAB, OSU, UMASS and UC Denver / DHHA. Matthew Wyatt was the presenter and the project targets allowing institutions to easily participate in a federated query cohort tool regardless of their underlying data structure and with minimal institutional infrastructure need. However, the FACE project is early in the design process, and discussions are ongoing about the caGrid/TRIAD component and how they will be utilized. Gopu Shrestha attended this teleconference to learn more about government.

12/19/2011: popHealth Webinar New Features in popHealth
Gopu Shrestha attended this webinar about using Drools to Analysis the measure and retrieve data from RDMS database. PopHealth is currently working with Ruby on Rails and MongoDB and they don't have any plan to work with Drools and RDMS database.

Partnerships / Relationships / Alliances (New in 2011)

a) Alessandro Moschitti, PhD, University of Trento, Italy: Relation extraction.
b) Clinical Information Modeling Initiative (CIMI). International consensus group with Detailed Clinical Models.
c) Consortium for Healthcare Informatics Research (CHIR), Matt Samore.
d) i2b2, Boston – Scrubbing patient data for in-house use. Multi-scrubber deployed and used. Integration of cTAKES into the Text Cell.
e) James Pustejovsky, PhD, Brandeis University, ISO TimeML.
f) Keith Toussaint, Business Development Manager, Amazon Web Services, expert consultant at Security Roundtable.
g) Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC) Leonard D'Avolio, PhD.
h) Multi-source Integrated Platform for Answering Clinical Questions (MiPACQ; funded by the NLM, NLM RC1LM01060).
i) National Quality Forum / MAT User Group: Jyoti Pathak, Dingcheng Li and Gopu Shrestha are working internally to come out with findings to share with MAT.

j) Open Health Natural Language Processing (OHNLP) outreach/Vanderbilt, Kaiser, and UMN/coordinating and leading the discussion.
k) Partnership with UCSD's NCBC iDASH. Purpose: create a web-based environment for distributing to the public common models, annotation schemes, and web services for tools and resources developed in the SHARP Area 4 grant.
l) Richard Wolniewicz, PhD, Director, NLP Advanced Technology, 3M Health Information Systems, expert consultant at Security Roundtable.
m) Sacsha Dublin, MD, Seattle Group Health: we are engaged in a project with Dr. Dublin to study the use of NLP technologies in the detection of pneumonia. It overlaps one of the SHARPn phenotyping projects.
n) SHARPc, Project 2B: We have been working with the SHARPc project on aspects of clinical decision support and discussing the use of CEMs and other SHARPn artifacts as a part of this project.
o) Southeast Minnesota Beacon Community. SHARPn is currently collaborating with this Beacon community for data normalization, NLP and Phenotyping.
p) Suresh Manandhar, PhD, University of York, UK: unsupervised machine learning and word sense disambiguation.
q) National Library of Medicine, UMLS team, embedding UMLS in cTAKES.

Operational Activities

a) SHARPn program organization is implemented with fostered social connections across projects. Individual project efforts synergized with timelines in synch; use cases vetted and determined for the first six months of focus.
b) Face-to-face collaboration has been fosterd in both intra-SHARP and cross-SHARP programs in cross-knowledge pollination and collaboration activities.
c) Project managers are responsible for day-to-day management, execution, and delivery of project team deliverables. Measures are monitored and documented as achievement of milestones by target dates and accomplishment of tasks in accordance with defined expectations. The project managers track progress (scope, resources and costs), proactively manage risk, track lessons learned and report to the stakeholders.
d) IRBs have been submitted and approved at all applicable sites.
e) Data Sharing issues have been raised with best practice sharing and inventory of existing agreements between institutions reviewed.
f) A cross-SHARP program synergy assessment was conducted with cross-SHARP area tasks mapped and plans for resourcing scoped.
g) Cross-SHARP collaborations have been identified and are being actively pursued; project aims & progress will be reported in next semi-annual report:
    i) SHARPs (Illinois) – phenotype implications of accessed data; address security in data segmentation and de-identification; ontology for policy; NLP to identify sensitive information in free text.
    ii) SMART Platform (Harvard) – common data model; access to clinical data – sandbox for developers.
    iii) SMART Platform (Harvard) & MD PnP – prototype communication between devices – data modeling standards.

2) Personnel / Hiring (ARRA Report)

Budgeted Personnel have remained consistent with justification approved.

Calendar Year / Quarter: 2011 / 1
Number of Jobs Count: 14.3

Calendar Year / Quarter: 2011 / 2
Number of Jobs Count: 13.5

Calendar Year / Quarter: 2011 / 3
Number of Jobs Count: 13.5

3) Grants Management (ARRA Report)

Expenditures have remained consistent with work scope approved.

Calendar Year / Quarter: 2011 / 1
Total Federal Amount of ARRA Expenditure: $1,649,397

Calendar Year / Quarter: 2011 / 2
Total Federal Amount of ARRA Expenditure: $2,346,573

Calendar Year / Quarter: 2011 / 3
Total Federal Amount of ARRA Expenditure: $3,467,203