Office of the National Coordinator for Health Information Technology

Strategic Health IT Advanced Research Projects (SHARP)

Progress Report

Reporting period: 1/1/2011 – 6/30/2011

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®

Collaborators:

- Agilex Technologies
- CDISC (Clinical Data Interchange Standards Consortium)
- Centerphase Solutions
- Deloitte
- Group Health, Seattle
- IBM Watson Research Labs
- University of Utah
- Harvard Univ. & i2b2
- Intermountain Healthcare
- Mayo Clinic
- Minnesota HIE (MNHIE)
- MIT and i2b2
- SUNY and i2b2
- University of Pittsburgh
- University of Colorado

1) 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 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 6) Evaluation Frameworks. The first two projects align with our Data Normalization theme, with Phenotype Applications and Performance Optimization span themes 1 and 2 (Normalization and Phenotyping); while the last two projects correspond to our third theme.

2) 2011 Progress Report - Executive Summary

Q1: The SHARP Area 4 project Team happy to promote the IBM Jeopardy Challenge. Out of the IBM TJ Watson Research Lab, represents a major advance in what is called "Deep Question Answering technology" the ability of a computer to do something that's far more challenging than chess: to understand natural human speech about a limitless range of topics, and to make informed judgments about them. Watson expects the science to elevate computer intelligence; to take human-to-computer communication to new levels; and to help extend the power of advanced analytics to make sense of vast quantities of structured and unstructured data. Watson on Healthcare For example, the Deep QA technology could, in the not so distant future, provide critical, timely information to physicians to help diagnose and treat patients. In SHARP Area 4 the UIMA-AS technology framework will be the backbone of the various tools and architecture. Initial practice use cases will be piloted in the Southeast MN Beacon. Members of the SHARP team attended HIMSS 11 as both exhibitor and presenters showcasing early technologies/research and establishing additional collaborations. The cNLP team released a second annotator. This release includes a Smoking Status Classifier that processes clinical documents and identifies patients' smoking status at the patient level as well as the document level. This pipeline will generate one of five smoking status categories: Past smoker, Current smoker, Smoker, Non-smoker, and Unknown.

Q2: The SHARP Area 4 (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 software developed by SHARPn will allow researchers and clinicians to pull together very different types of healthcare data and ask questions about disease, prevention, outcomes, and care delivery. Answering these questions requires a robust, high-quality dataset, which can be generated by software in the SHARPn cloud-computing environment. SHARPn hosted a roundtable for cloud-deployed clinical Natural Language Processing May 22-24 in Seattle. The roundtable brought 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 ("the cloud"). SHARPn also ran a “tracer-shot” pilot of its UIMA Pipeline; sending de-id patients through its UIMA pipeline from Intermountain Healthcare and Mayo Clinic processing all of the stages of data normalization, clinical natural language processing, persisted in MySQL database and utilized a Drools based phenotyping process. The SHARP Area 4 held its face-to-face on June 30-July1. Eighty-nine experts from dozens of research institutions and the federal government attended the conference at the University of Minnesota Rochester. Areas of specialization ranged from informatics and computer science to medicine and administration.

SHARP Area 4 Announcements can be found at the following URL:
http://informatics.mayo.edu/sharp/index.php/Main_Page#Announcements

3) Research Update: Cross-integrated suite of project and products

a) Clinical Data Normalization & Evaluation Framework

In 2010, the two defined projects of Clinical Data Normalization and Evaluation Framework that represent the 'bookends' of the program were combined for scope synergies and resource sharing.

i) *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 NLP where appropriate, against normalized forms, and tabulate discordance. Use cohort identification algorithms in both EMR data and EDW data. (normalize against CEMs).

ii) *Progress*:

(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) Formalized Meaningful Use vocabularies for LexGrid server

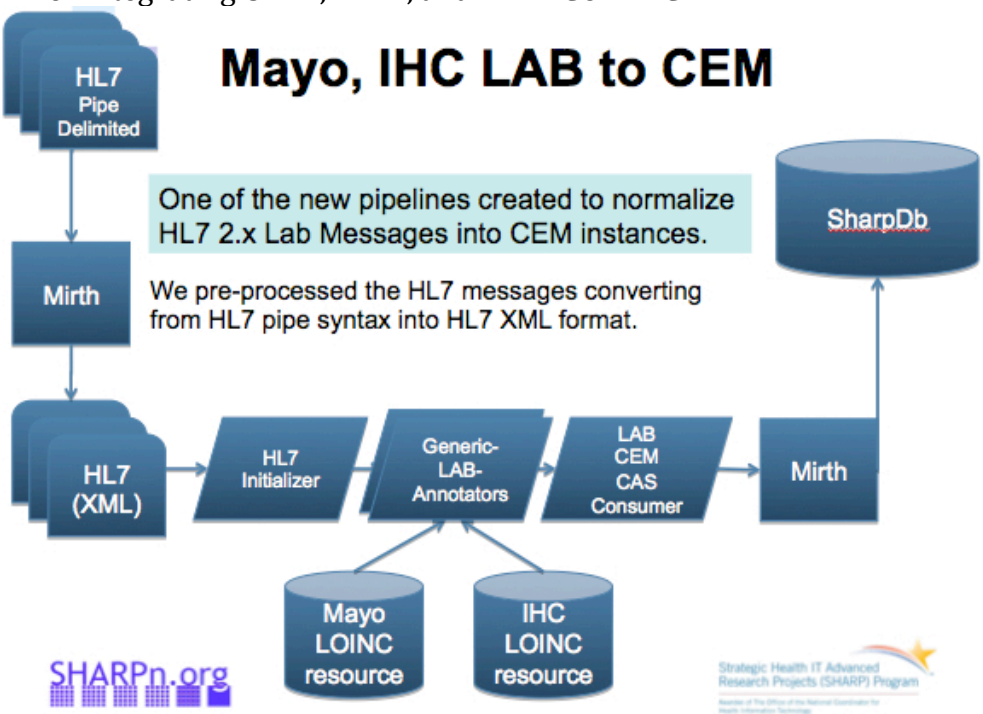
(3) Designed other components of Data Normalization framework (Terminology Services - NHIN connections).

(4) Maintained collaboration with Area 3 with CEM models and persistent layer cohesion.

(5) Identified missing CEMs for data (and classes of data) in use-cases.

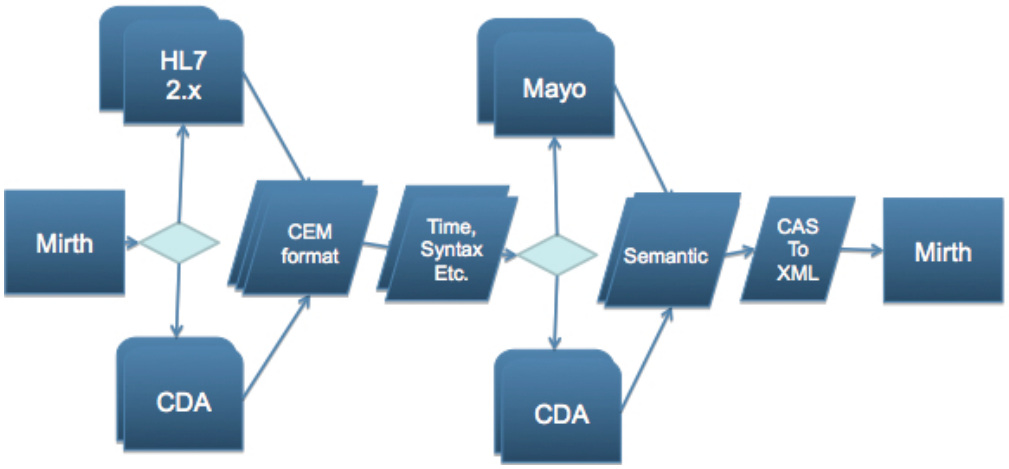
iii) *Milestones Reached*:

(1) 'Tracer-Shot' architectural plan executed (Step-based knowledge acquisition) for integrating UIMA, BPEL, and NHIN CONNECT



(2) Implemented CONNECT software environment (C32 specs; CEM subsets formulate XML docs (part of meaningful use).
(3) Persistence Channels defined
  (a) One Channel per model
  (b) Data stored as an XML Instance of the model
  (c) Fields extracted from XML to use as indices
  (d) XML Schema defined for each model
  (e) Stored using database transactions

iv) *Next Steps*:
(1) Single CEM for multiple OBX segments
(2) Efficiently utilize terminology services
(3) Incorporate a library for HL7 clean-up routines
(4) Increase scope of vocabulary standardization
(5) Enhancements for the Drug Annotator
(6) Context enhancement issue
(7) Drug name surprises
(8) Review sources used for normalization opportunities E.g.
  (a) In HL7 OBR Segments
    (i) Standardize Service ID (Codes)
  (b) In HL7 OBX Segments
    (i) Standardize Units
    (ii) Standardize Reference Ranges
    (iii) Standardize Normal Flags



b) Clinical Natural Language Processing (cNLP)

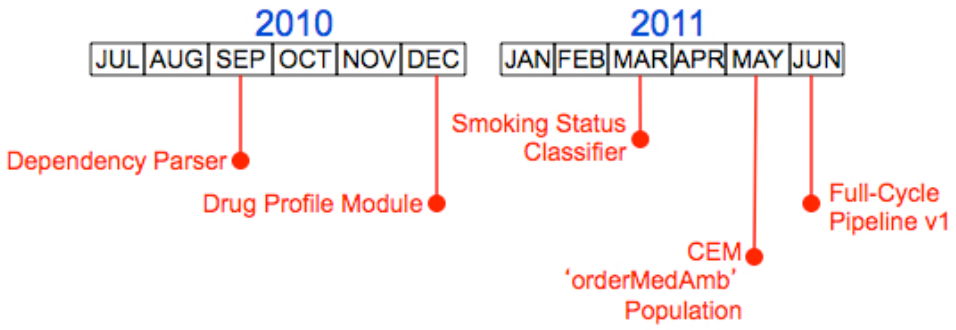
i) *Aims*: Information extraction (IE): transformation of unstructured text into structured representations and merging clinical data extracted from free text with structured data.

ii) *Progress*:

(1) The Common Type System project has produced the schema for a common UIMA type system. This type system gives a clear place for semantics that usable for high-throughput phenotyping.
(2) Developing a library of integrated de-identification systems with surrogate generation (MIT/SUNY and MITRE)
(3) Coreference resolution module is under development
(4) Relation extraction and active learning
(5) Seed corpus generation and its deidentification
(6) Annotation schema development based on Clinical Element Model
(7) Annotation guidelines development for (6) and pilot annotations
(8) Stratified corpus methodology
(9) Identification of a set of eventive UMLS entity types and relations for the automatic extraction of medical events
(10) Development of evaluation workbench
(11) Aligning annotations with ISO Linguistic standards and other conventions
(12) Discussions with the NLM team about licensing the dictionaries released with cTAKES
(13) Phenotype rules implemented in GELLO

iii) *Milestones Reached*:

(1) Type system for the NLP project (May, 2011)
(2) MIT/SUNY v1 of deidentifier and surrogate generator 1.0 release
(3) 150K token seed corpus generation and de-identification
(4) Annotation model and schema based on CEMs for diseases/disorders, signs/symptoms, medications, anatomical sites, procedures, labs
(5) Annotation guidelines finalized
(6) CEM OrderMedPopulation, May 2011
(7) Full-cycle pipeline (free text-NLP-CEM) as part of the SHARPn v1 pipeline
  (a) All medications in Mayo dataset extracted with cTAKES (NLP method)
  (b) Processed 360,452 notes for 10,000 patients
  (c) 3,442,000 CEMs were created
  (d) Processing time: 1.6 sec/doc



(8) Security roundtable meeting in Seattle: A set of recommendations for the novel legal and governance issues regarding the proper stewardship and use of clinical data
(9) 450 clinical notes have been automatically part of speech annotated and manually corrected, and automatically syntactically parsed; 25% of those trees have been manually corrected, including adding traces and function tags.
(10) Incorporation of tree kernel tools
(11) Early beta draft of the evaluation workbench
(12) Smoking status module to cTAKES released, March 2011
(13) Early beta version of the coreference module
(14) Collaborated with the HTP team on GELLO implementation

iv) *Next Steps*:
(1) Full integration of the common type system into cTAKES
(2) Improvement of the de-identification tools and integration with the common type system
(3) Selection and de-identification of the stratified corpus.
(4) Assisting MITRE in development of de-identification annotations
(5) Preparation of a document summarizing lessons and recommendations from the cloud security roundtable.
(6) Continued gold standard annotation
(7) Retraining of tools based on the gold standard
(8) Development of UMLS Relation Extraction system
(9) Release of Evaluation workbench v1
(10) Complete CEM management, viewing, translation, download, and requesting facility
(11) Test MIST de-identification on SHARP Pilot Corpus and i2b2 Challenge Evaluation data to evaluate MIST's ability to preserve coreference chains and transformation from date-surrogate to date-original forms.
(12) Integration of UIMA-wrapped version of MITRE's negation/uncertainty module into cTAKES
(13) Release of coreference module
(14) Release of side effects module
(15) Software development group (SDG) for rapid component integration

c) High throughput Phenotyping (HTP)

i) *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.

ii) Clinical Element Model identification and creation for phenotyping algorithms
(1) Members: Cui Tao, Susan Welch, Tom Oniki, Craig Parker
(2) Overarching goal: This project under HTP is focusing on identification and modification of existing, and where applicable, creation of new CEMs (Clinical Element Models) for Electronic Health Record (EHR) based phenotyping algorithms.
(3) Progress till date:

(a) Identification of existing, and relevant, CEMs for Peripheral Arterial Disease (PAD), Type 2 Diabetes (T2D), Hypothyroidism, and Community Acquired Pneumonia.
(b) Several communications and meetings with SHARP Area 4 Data Normalization team to modify existing CEMs, or create new CEMs, where necessary.
(c) Categorization of CEMs (phenotype-independent Vs. phenotype-specific) to facilitate future implementation of the CEM browser.
(d) Resource Description Framework (RDF) based representation of CEMs.

(4) Milestones reached:
(a) Identification and documentation of CEMs for 3 different phenotyping algorithms (PAD, T2D, Hypothyroidism) in the SHARP Area 4 wiki.
(b) Preliminary analysis of "gap" between existing CEMs, and what is required for the phenotyping algorithms

(5) Directional inflections: None

(6) Next steps and milestones:
(a) Preliminary representation of RDF-based CEMs
(b) Collaboration with data normalization team for representation of patient-specific instance data conformant to the selected CEMs for PAD, T2D, and Hypothyroidism.
(c) Manuscript publication/white paper on CEM analysis.

iii) RDF representation of Clinical Element Model

(1) Members: Cui Tao, Craig Parker, Tom Oniki

(2) Overarching goal: This project is a collaborative project with the normalization team. It focuses on a formal definition of the CEM using semantic web specifics.

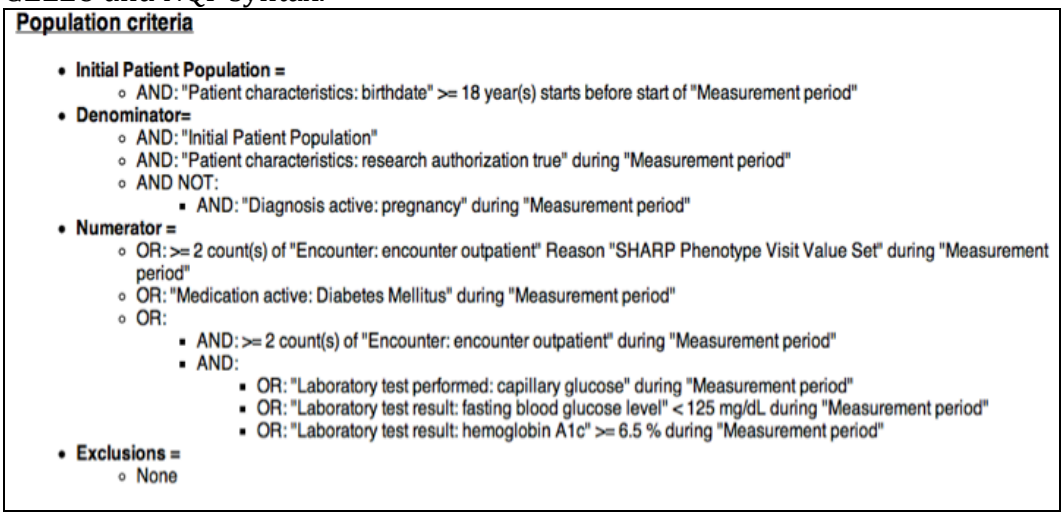
(3) Progress till date:
(a) Identifying representative models from the CEM repository at Intermountain Healthcare
(b) Several communications and meetings with SHARP Area 4 Data Normalization team to discuss the feasibility of the RDF representation

(4) Milestones reached:
(a) Identification and documentation of CEM use cases in the CEM library.
(b) Preliminary analysis of the benefits on RDF representation of CEM
(c) Manuscript of RDF-based representation of CEM accepted for publication at the 2011 AMIA Fall Symposium, Washington DC.

(5) Directional inflections: None

(6) Next steps and milestones:
(a) Preliminary representation of RDF-based CEMs and load them in a RDF triple store
(b) Collaboration with data normalization team for the semantic definition of CEM using semantic web notations

iv) Structured protocol representation and phenotyping logic
  (1) Members: Jyoti Pathak, Herman Post, Darin Wilcox, Jeffrey Ferraro, Mark Arratoon, Landen Bain, Mike Conway (UCSD)
  (2) Overarching goal: This project under HTP is focusing on leveraging the standards-based protocol representation model for structured modeling of phenotyping algorithms.
  (3) Progress till date:
    (a) Analysis of 13 eMERGE (Electronic Medical Records and Genomics) network phenotyping algorithms for identifying commonalities, differences, data elements used, phenotyping logic used etc.
    (b) Analysis of CDISC protocol representation model for the purposes of the SHARP HTP project
    (c) Analysis of other protocol representation specifications, including Drools, GELLO and NQF syntax.



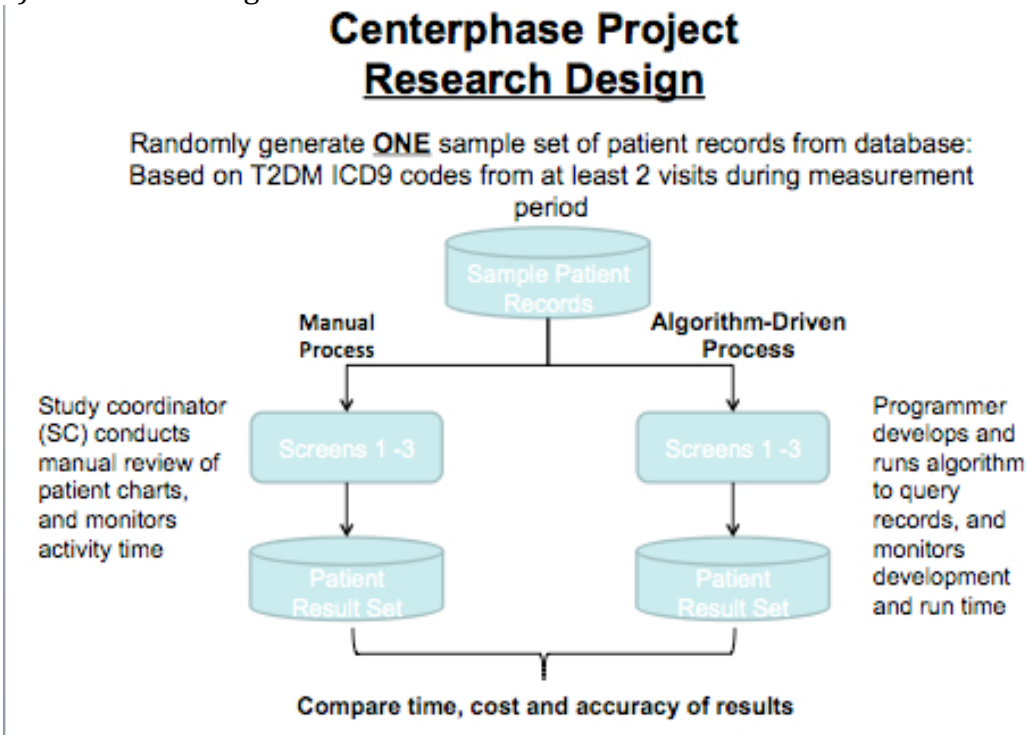
Population criteria

- **Initial Patient Population =**
  - AND: "Patient characteristics: birthdate" >= 18 year(s) starts before start of "Measurement period"
- **Denominator=**
  - AND: "Initial Patient Population"
  - AND: "Patient characteristics: research authorization true" during "Measurement period"
  - AND NOT:
    - AND: "Diagnosis active: pregnancy" during "Measurement period"
- **Numerator =**
  - OR: >= 2 count(s) of "Encounter: encounter outpatient" Reason "SHARP Phenotype Visit Value Set" during "Measurement period"
  - OR: "Medication active: Diabetes Mellitus" during "Measurement period"
  - OR:
    - AND: >= 2 count(s) of "Encounter: encounter outpatient" during "Measurement period"
    - AND:
      - OR: "Laboratory test performed: capillary glucose" during "Measurement period"
      - OR: "Laboratory test result: fasting blood glucose level" < 125 mg/dL during "Measurement period"
      - OR: "Laboratory test result: hemoglobin A1c" >= 6.5 % during "Measurement period"
- **Exclusions =**
  - None

  (4) Milestones reached:
    (a) Evaluation results available for 13 eMERGE algorithms with respect to data elements used, terminologies used, and phenotyping logic for representation of the algorithms. Manuscript accepted for publication at the 2001 AMIA Fall Symposium, Washington DC. for submission to the 2011 AMIA Annual Symposium is available.
    (b) Preliminary representation of the Medicinally managed diabetes algorithm using Drools.
  (5) Directional inflections: None
  (6) Next steps and milestones:
    (a) Continue evaluation of Drools and NQF for syntactic representation of phenotyping algorithms.
    (b) Library of phenotyping algorithms to identify cohorts of patients with diseases and conditions of interest for clinical trials, quality improvement, disease registry, CMS etc.

v) Phenotyping and data quality (section H)
   d) Members: Kent Bailey, Susan Welch, Susan Fenton
   e) Overarching goal: This project under HTP addresses issues relevant to variation and heterogeneity of EHR data across institutional boundaries for any given disease or phenotype.
   f) Progress:
      i) Development of a study design for EHR data comparison and analysis at 2 different academic medical centers: Mayo Clinic and Intermountain Healthcare/University of Utah.
      ii) Joint IRB under submission and review.
   g) Milestones reached:
      i) Study design for EHR data comparison across 2 different medical institutions for PAD, T2D, and Hypothyroidism.
   h) Directional inflections: None
   i) Next steps and milestones:
      i) Evaluation of variation of EHR data for T2D at Mayo Clinic an Intermountain Healthcare/University of Utah.
      ii) Manuscript publication/white paper on data variation and heterogeneity.

4) Market value and financial aspects of EHR-derived phenotyping
   a) Members: Jeff Tarlowe, Gary Lubin
   b) Overarching goal: This project under HTP is studying the financial aspects of manual phenotyping vs. electronic phenotyping using the proposed SHARPn tools and technologies.
   c) Research Design



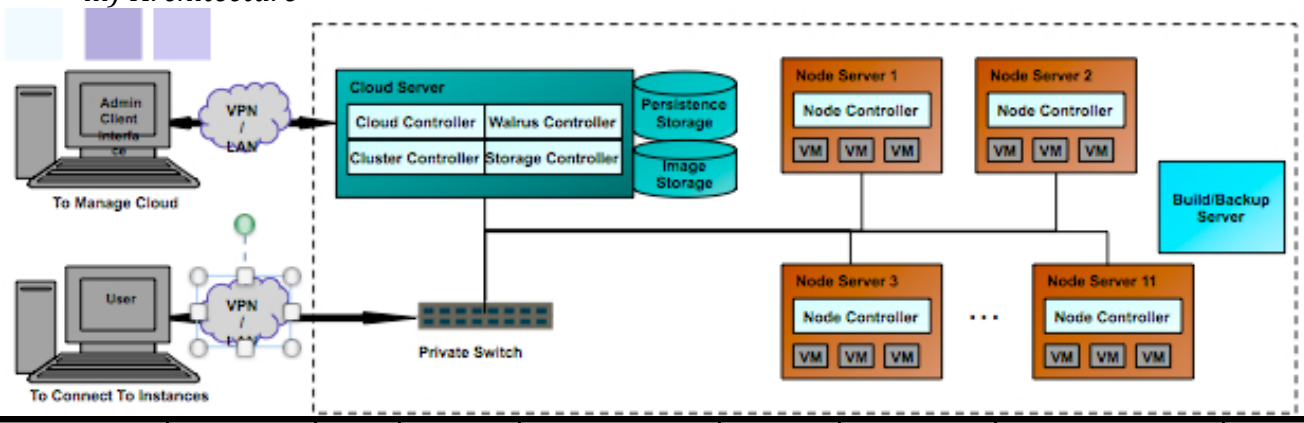
d) Progress:
   i) Design of approach to determine market metrics for EHR derived phenotyping algorithms
   ii) Development of use cases and planning for comparing manual Vs. EHR-based clinical phenotyping, and understanding business and financial implications.
e) Milestones reached:
   i) Project plan to test market value of new EHR based algorithms for cohort identification to facilitate clinical trials.
   ii) Identification of market metrics for cohort identification based on false positives, and false negatives.
   iii) Execution and evaluation of the approach for testing market value of new EHR based algorithms for cohort identification to facilitate clinical trials. This evaluation was done only for 50 subjects.
f) Directional inflections: None.
g) Next steps and milestones:
   i) Continue the execution and evaluation of approaches for larger cohort of 500 subjects.
   ii) Manuscript publication/white paper determining economic value and financial aspects for EHR derived phenotyping.

h) Infrastructure & Scalability

i) *Aims*: Consult on pipe line design / architectures / configuration; Work with team leads to identify "fit" (or not) of UIMA into subprojects; Develop and deploy virtual machine images that can dynamically scale in cloud computing environments.

ii) *Progress*:
(1) Installation of informatics "SHARP" Cloud system at Mayo
(2) Mirth Enhancements
   (a) Implemented NwHIN XDR connector capability
   (b) Implemented UIMA connector capability
   (c) Created NwHIN Aurion XDR adapter
(3) Channels Created
(4) Installation and configuration of tools on IHC side and SHARP Cloud
(5) 30 de-id IHC patients through the UMIA pipeline
   (a) Resulting in 134 Thousand CEMS generated
(6) Processing of 10,000 patients Meds, Labs, Billing data
   (a) Resulting in 15 Million CEMS generated
(7) Persisted from CEM to MySQL database
(8) Drools Phenotype processing on IHC dataset
   (a) 30 de-identified IHC patients
   (b) Resulting in 134 Thousand CEMS generated

iii) Architecture



Hardware	No. of Physical Machines	CPU	Memory	Disk	Disk Space	Networking	Functionality	No. of NICs
Cloud Server	1	8	12 GB	10000 RPM SAS	1 TB	1 Gbps	Cloud, Walrus, Cluster and Storage Controller	4
Node Server	1	8	32 GB	10000 RPM SAS	1 TB	1 Gbps	Node Controller	4
Node Server	8	24	128 GB	10000 RPM SAS	600 GB/600 GB	1 Gbps	Node Controller	4
Node Server	1	8	64 GB	7200 RPM SATA	1 TB/1 TB	1 Gbps	Node Controller	4
Node Server	1	8	32 GB	10000 RPM SAS	4 TB	1 Gbps	Node Controller	4
Build/Backup Server	1	2	8 GB	7200 RPM SATA	2 TB	1 Gbps	Build and Backup	2
Storage	2			10000 RPM SAS	7.5 TB	1 Gbps	Persistence and Image Storage	
Storage	2			10000 RPM SAS	3.6 TB	1 Gbps	Volume Storage	
Cisco 48 Port Switch	2		1 GB					

iv) *Milestones Reached*:

(1) Cloud Infrastructure Installed
(2) NwHIN Communications Verified between IHC & Mayo
(3) UIMA pipeline operational on Cloud
(4) Clinical Element Model Instance stored in MySQL Database
(5) Drools operational on Cloud

v) *Next Steps*:
(1) Single CEM for multiple OBX segments
(2) Efficiently utilize terminology services
(3) Incorporate a library for HL7 clean-up routines
(4) Increase scope of vocabulary standardization

i) Data Quality

i) *Overview*: 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.

ii) *Progress*:
(1) Actively participated with Area 4 project areas (Data Normalization, HTP, and cNLP) meetings
(2) Finalization of protocol for inter-institutional comparison of data relevant to Type 2 DM phenotype
(3) *Specifications for mapping 3 eMerge Phenotyping Algorithms from Clinical Element Model to the algorithm rules' abstracted EHR input descriptions.*
(4) *2) Generalization and categorization of 3 eMerge Phenotyping Algorithm rules' abstracted EHR input descriptions.*
(5) Initial plan for de-identification and sharing of data
(6) IRB approved protocol(s) to retrieve potential DM cases at Mayo and to access data for study design at Intermountain Health Care
(7) Initial retrieval of Mayo cohort
(8) Preliminary analysis of Mayo cohort
(9) Highly Preliminary comparison of cohorts
(10) Execution of Centerphase study to assess cost-benefit comparison between electronic and manual review for flagging high-risk Type 2 DM.

iii) *Milestones Reached*:
(1) A common protocol for inter-institution data comparison has been developed.
(2) An initial Mayo dataset is nearing completion
(3) A preliminary look at IHC denominator for the study, and preliminary comparisons
(4) Integration of Centerphase efforts and DQ project
(5) Close working relationship of Mayo, IHC, Centerphase Data groups

iv) *Next Steps*:
(1) Extend the mapping specifications and generalizations developed above to cover more eMerge Phenotyping Algorithms as well as algorithms to identify the reportable cohorts for AHRQ quality measures and cohorts potentially eligible for clinical trials.
(2) Extend the Medicinally Managed Diabetes Phenotype Algorithm secondary usage use case beyond algorithm execution. Define requirements for data analysis to

support data quality, practice organization variations in EHR data and other usages of the extracted and transformed data sets generated from source clinical elements and prepared as inputs that meet the clinical algorithms' abstracted EHR input descriptions.

(3) Run Association Mining step on initial Mayo, IHC datasets
(4) Use results to update retrieval protocols
(5) Carry out comparison of aggregate data, to identify potential differences in data collection and interpretation
(6) Create framework for human verification of data. Use grant resources to sample charts at each institution to verify/ disconfirm results of electronic pull / cTAKES results.
(7) Develop plan to propagate DQ/H results to phenotype algorithm results.
(8) Continue with Centerphase comparative analysis of manual/electronic processing of DM cases.
(9) Manuscript publication/white paper on data variation and heterogeneity.
(10) Replication study design with University of Texas collaboration.

5) Program Outputs

a) Products

i) CEM Database Design – 6 tables

ii) MIRTH Channels
(1) Sample XDR Channel – to push data via NwHIN Gateway
(2) ReceiveXDRMessage – to receive data via NwHIN Gateway
(3) CemAdminDxtoDatabase – Store Billing Codes to CEM Database
(4) CemLabToDatabase – Store Labs Results to CEM Database
(5) CemMedicationToDatabase – Store Meds to CEM Database

iii) NLP
(1) UML representing the Common Type system and preliminary XML representation
(2) MIT/SUNY deidentification tool as part of the SHARPn library
(3) Draft document summarizing lessons and recommendations from the cloud security roundtable
(4) Annotated data
(5) Early version of Evaluation workbench
(6) Smoking status module for cTAKES, March 2011
(7) CEM OrderMedPopulation, May 2011
(8) End-to-end pipeline, v1, June 2011

b) Publications and Presentations

i) Recent/accepted/published

(1) 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).

(2) 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).

(3) 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).

(4) Conway MA, Pathak J. Analyzing the Prevalence of Hedges in Electronic Health Record Oriented Phenotyping Algorithms. AMIA 2011 (poster).

(5) –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).

(6) Submission to MIX-HS 2011 on Common Type System

(7) Dmitriy Dligach and Martha Palmer. 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.

(8) Dmitriy Dligach and Martha Palmer. 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.

(9) Jinho Choi and Martha Palmer, 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.

(10) Jinho Choi and Martha Palmer: 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

(11) Savova GK, Chapman WW, Elhadad N and Palmer M. 2011. Shared annotated resources for the clinical domain. AMIA ann symp. Panel.

(12) Chute CG, Pathak J, Savova GK, Bailey K, Schor M, Hart L, Beebe C and Huff S. 2011. The SHARPn project on secondary use of electronic medical record data: progress, plans an possibilities. AMIA ann symp.

(13) Evaluation as a driver in Software Communities, Lynette Hirschman, 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

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

(15) Recent efforts in clinical NLP: Uncertainty discovery through NLP, Cheryl Clark, presentation to Natural Language Processing Workshop, i2b2 Academic Users Group, Boston, June 28, 2011
(16) MITRE System for Clinical Assertion Status Classification, JAMIA 2011;Published Online First: 22 April 2011 doi:10.1136/amiajnl-2011-000164

i) Planned
(1) Manuscript/s on MIT/SUNY deidentifier and surrogate generator
(2) Manuscript/s on coreference resolution for clinical text
(3) Lessons and recommendations from the cloud security roundtable
(4) Manuscript/s on Evaluation workbench
(5) Manuscript/s on CEMs and NLP
(6) Manuscript/s on relation extraction
(7) Manuscript/s on active learning methodology
(8) Manuscript/s on comparison of coding, lab, and med data between institutions
(9) Manuscript/s on cost-benefit analysis of computer-assisted screening of HER for phenotype identification or high-risk status identification
(10) Manuscript/s specifying mapping of CEMs to Use Case algorithms' input specifications: focusing AMIA Clinical Trials or Translational Informatics conference.
(11) Manuscript for submission to JBI publication focusing application and use of standards with input/authorship from among the SHARPN team.

6) Events

December 13-15, 2010: 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. Szolovits gave a talk at Basis Technology, Cambridge MA, on January 4.

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. Szolovits lectured in the Distinguised Lecture series at Wayne State University, Detroit, on March 29.
4/5-4/7/2011 – 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 Chris Chute, Guergana Savova, Wendy Chapman, Lynette Hirschman, David Carrell. Our attendees: Chris 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. Szolovits gave a keynote address at the Workshop on Learning from Clinical Text at the International Conference on Machine Learning, Seattle, July 2, 2011.

7/11/2011-7/12/2011 SHARPFest: Washington, DC
Attendees: C.Chute, S.Huff, L.Hart, J.Pathak, G.Savova

7) Partnerships / Relationships / Alliances (New in 2011)

a) I2b2, Boston -- Scrubbing patient data for in-house use. Multi-scrubber deployed and used. Integration of cTAKES into the Text Cell
b) Richard Wolniewicz, PhD, Director, NLP Advanced Technology, 3M Health Information Systems, expert consultant at Security Roundtable
c) Keith Toussaint, Business Development Manager, Amazon Web Services, expert consultant at Security Roundtable.
d) Multi-source Integrated Platform for Answering Clinical Questions (MiPACQ; funded by the NLM, NLM RC1LM01060).
e) Alessandro Moschitti, PhD, University of Trento, Italy: relation extraction
f) Suresh Manandhar, PhD, University of York, UK: unsupervised machine learning and word sense disambiguation
g) James Pustejovsky, PhD, Brandeis University, ISO TimeML

h) 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.
i) 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.
j) Sacsha Dublin, MD, Group Health, Seattle: 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.
k) Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC) Leonard D'Avolio, PhD.
l) Consortium for Healthcare Informatics Research (CHIR), Matt Samore

8) 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 have been fosterd in both intra-SHARP and cross-SHARP program's 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; best practices in the data normalization & security audits of the pipeline; address security in data segmentation and de-identification; ontology for policy; NLP to identify sensitive information in free text.
    ii) SMART (Harvard) – common data model; access to clinical data – sandbox for developers.
    iii) SMART (Harvard) & MD PnP – prototype communication between devices – data modeling standards.

9) Personnel / Hiring (ARRA Report)

Budgeted Personnel have remained consistent with justification approved. No significant changes forecasted for Q3/4 of 2011.

Calendar Year / Quarter: 2011/ 1
Number of Jobs Count: 14.3
Description of Jobs Created: Investigators, Project Managers, Research Associates, Senior Analyst/Programmers, Analyst/Programmer, Project Assistant,Chief Executive Officer, Chief Operating Officer, Principal Investigator, Co-Investigator, Research Software Developer, Data Manager I, Data Manager II, Investigator, Graduate Student, Post Doctorate, Research Assistant, Research Project Assistant, Research Support Specialist, Research Aide, Senior Research Aide, Fellow, Project Manager, Programmer Analyst, Research Specialists, Research Staff Member, Key Personnel, Project Director, Software Engineer-Cnslt, Medical Informaticist-Cnslt, Chief Medical Info Officer, Medical Vocab Engineer-Sr, Medical Informaticist Sr, Medical Informaticist-Cnslt, Data Architech-Sr, Medical Vocab Engineer-Sr, Senior Manager, Senior Consultant, HIE/Terminology Solution Architect

Calendar Year / Quarter: 2011 / 2
Number of Jobs Count: 13.5
Description of Jobs Created: Investigators, Project Managers, Research Associates, Senior Analyst/Programmers, Analyst/Programmer, Project Assistant,Chief Executive Officer, Chief Operating Officer, Principal Investigator, Co-Investigator, Research Software Developer, Data Manager I, Data Manager II, Investigator, Graduate Student, Post Doctorate, Research Assistant, Research Project Assistant, Research Support Specialist, Research Aide, Senior Research Aide, Fellow, Project Manager, Programmer Analyst, Research Specialists, Research Staff Member, Key Personnel, Project Director, Software Engineer-Cnslt, Medical Informaticist-Cnslt, Chief Medical Info Officer, Medical Vocab Engineer-Sr, Medical Informaticist Sr, Medical Informaticist-Cnslt, Data Architech-Sr, Medical Vocab Engineer-Sr, Senior Manager, Senior Consultant, HIE/Terminology Solution Architect

10) Grants Management (ARRA Report)

Expenditures have remained consistent with work scope approved. No significant changes forecasted for Q3/4 of 2011.

Calendar Year / Quarter: 2011 / 1
Total Federal Amount of ARRA Expenditure: $1,649,397.07

Calendar Year / Quarter: 2011 / 2
Total Federal Amount of ARRA Expenditure: $2,346,573.17