Office of the National Coordinator for Health Information Technology

Strategic Health IT Advanced Research Projects (SHARP)

Semi-Annual Progress Report

Reporting period: 1/1/2012 – 6/30/2012

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

Collaborators:

- Agilex Technologies
- CDISC (Clinical Data Interchange Standards Consortium)
- Deloitte
- Group Health, Seattle
- IBM Watson Research Labs
- University of Utah
- Harvard Univ. & i2b2
- Intermountain Healthcare
- Mayo Clinic
- Mirth Corporation, INC.
- MIT and i2b2
- MITRE Corp.
- Regenstrief Institute, Inc. SUNY and i2b2
- University of Pittsburgh
- University of Colorado
- University of California, San Diego

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) 2012 Progress Report - Executive Summary

Q1 (ARRA Report): During Q1 of 2012, the SHARPn program has progressed with its open-source tools addressing clinical barriers to the broad-based, facile, and ethical use of EHR data for secondary purposes. In an effort to accelerate SHARPn deployment of methods & tools into practice, the program will broaden its sub-award partnership to include Indiana University in affiliation with Regenstrief Institute and the Indiana HIE as well as fostering partnerships with Beacon communities. The SHARPn program has upheld six strongly intertwined, mutually dependent projects, that constitute an ensemble of related research and development; however, a deliberate shift has been made in the program to define efforts and outcomes for both the research community and clinical practice consumers. Domain Progress: Q1 2012
A) International consensus group formed with Detailed Clinical Models, Clinical Information Modeling Initiative (CIMI) collaboration. B) Clinical Models for Laboratory & medication data released. C) Deployed Multiscrubber, a meta-classifier approach to de-identification in NLP. D) The SHARPn seed corpus consisting of 200K words of clinical narrative is fully annotated and adjudicated for treebank and propbank annotations. E) Completed pilot of Drools inference engine tied with UIMA pipeline in Cloud environment. F) Successfully modeled type II diabetes phenotype in PopHealth G) Finalized requirements and architecture for Phenoportal - development began in March. H) Succesfully drafted eMeasures using National Quality Form (NQF) Measure Authoring Tool (MAT) for type II diabetes phenotype. I) Successfully working with NQF on MAT tool usability issues. Actively providing support to Northwestern University eMERGE team and have modeled additional 10 eMERGE algorithms using the Quality Data Model (QDM) from NQF. J) SHARPn cloud computing environment evaluation by collaborators. K) Pipeline "tracer-shot" completed on Laboratory & Medication feeds.

Q2 (ARRA Report): SHARPn held its 3rd Annual Face-to-Face Summit. This year included over 125 attendees. The meeting included tutorials, published paper and poster presentation, technical demos and code-a-thon opportunities: http://informatics.mayo.edu/sharp/index.php/Annual_Gathering. Attendees learned more about SHARPn and the health IT field, explored technologies and tools that enable secondary uses of EHR data, participated in a hands-on tutorial from experts on how to better leverage secondary data using open source tools, presented their paper or poster on related topics through a peer reviewed process and worked side-by-side with community leaders to develop and refine customer and stakeholder requirements. In the NLP world, the cTAKES open-source software, version 2.5.0 was released and includes: a new attributes (assertion) extractor, a semantic role labeler, an additional sectionizer, updates to the coreference resolver, and an updated part-of-speech model. For the latest updates visit: www.sharpn.org

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

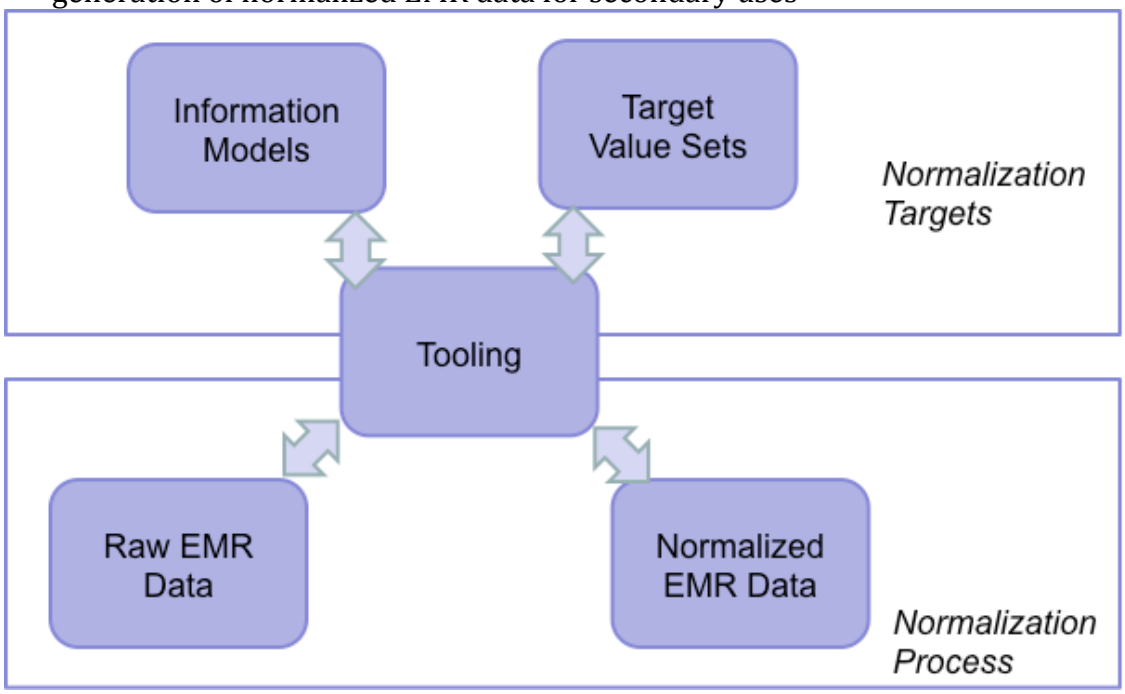
3) Research Update: Cross-integrated suite of project and products

a) Clinical Data Normalization & Evaluation Framework

i) Data Normalization
   (1) Comparable and consistent data are foundational to secondary use

ii) Clinical Data Models – Clinical Element Models (CEMS)
   (1) Basis for retaining computable meaning when data is exchanged between heterogeneous computer systems.
   (2) Basis for shared computable meaning when clinical data is referenced in decision support logic.

iii) *Aims:* To conduct the science for realizing semantic interoperability and integration of diverse data sources; To develop tools and resources enabling the generation of normalized EMR data for secondary uses



iv) *Modeling* In SHARPn, we've taken the models used at IHC and have tried to apply them to other settings
   (1) One concept is findings that include the attribute type (hair color) with the finding (brown)
   (2) Evaluation style - looking at a particular finding related to an attribute
   (3) Assertion style - look at a particular attribute of a person
   (4) Creates a different decomposition of the information
   (5) Typically use the evaluation style when the finding is a number (it would be strange to say (BP80) as a data result because you'd need many, many data results.
   (6) Both evaluation and assertion styles are accurate and unambiguous
   (7) Assertion styles allow each assertion to become a present/absent column for statistical analysis
   (8) Assertion styles are best for reasons, complications, final dx, etc.

(9) Conclusion: you need both.

v) *Modeling Accomplishments*:
   (1) CORE and Secondary Use Models/terminology patient, meds, labs, administrative dx and procedure, disease/disorder
      (a) CORENotedDrug -> SecondaryUseNotedDrug
      (b) COREStandardLab -> SecondaryUseStandardLab (+ 6 data type-specific models)
      (c) COREPatient -> SecondaryUsePatient
   (2) Definition of value sets in terms of standards where possible
   (3) CEM Browser
   (4) CEM Request Site: https://intermountainhealthcare.org/CEMrequests

vi) *Modeling Lessons Learned*
   (1) Open tools would be a great contribution to the interoperability. Examples:
      (a) mapping terminology, e.g., local codes to LOINC/HL7/SNOMED
      (b) mapping models, e.g., HL7 messages/CDA documents to CEMs, CEMs to ADL, etc.
      (c) generating sample instances
      (d) communicating information
         (i) browsers
         (ii) generating documentation
   (2) Documentation is essential – we need to produce more , despite communication barriers (verbally or in written word) the intricacies and complexities needed to make an effort like this work
   (3) "One model fits all" is difficult to achieve with multiple models out there.
      (a) Clinical Trials (e.g., CDISC CSHARE) vs Secondary Use (e.g., SHARPn)
      (b) Proprietary EMR (e.g., GE Qualibria) and Open Secondary Use (e.g., SHARPn)
      (c) value set differences
   (4) The root of all modeling questions: Precoordination vs. postcoordination and what to store in the model instance vs. leave in the terminology
      (a) Clinical drug vs. drug name/form/strength/ route
      (b) LOINC code vs lab test/method
      (c) Display names
      (d) Drug classes

vii) *Modeling Next Steps*:
   (1) Request site full launch
   (2) Browser launch
   (3) Browser enhancements (terminology integration, knowledge repository integration)
   (4) Better documentation
   (5) Model enhancements
   (6) Tooling – explore collaborations

viii) *Value Sets Accomplishments*:
   (1) Terminology value sets define the valid values used in the models
   (2) Terminology standards are used wherever possible
   (3) SHARP Value Sets published as CTS2 Resolved Value Sets

(a) ICD-9
(b) RxNorm
(c) LOINC
(d) ECIS
(e) SNOMED-CT
(4) CTS2 Services Created For
(a) SNOMED-CT Simple Refsets and Dynamic Value Set Definition
(b) ECIS Value Sets and Maps
(5) CTS2 Participation in CIMI effort

ix) *Value Sets Next Steps*:
(1) CTS2 Service for RxNorm and UMLS access
(a) Allow creation of new value sets
(b) Links to concept descriptions and relationships
(2) Integrate BioPortal Value Set Services
(a) Bridge Bioinformatics Ontology / Clinical Terminology
(3) Complete ECIS Server
(a) Will aid in CEM to ADL migration
(4) CTS2 Services available to CIMI project
(5) Integration
(a) Create Pipeline specific services coupled directly to CTS2
(b) Tighter coupling of Concept Domains and Value Sets
(c) Ajax Widgets for viewing, integrating and Authoring

b) Clinical Natural Language Processing (cNLP)

Overarching goal: general purpose reusable methodology and toolset for semantic text analytics of the clinical narrative applicable to a variety of use cases such as secondary use of the EMR, Meaningful use, Comparative effectiveness, Clinical investigation (e.g. patient cohort identification, phenotype extraction), Epidemiology, Clinical practice, Pharmacogenomics/Pharmacogenetics, clinical Question Answering to name a few.

i) 6 Core Clinical Element Model (CEM) templates as normalization targets for SHARP NLP
ii) The Science of NLP
(1) Research Areas
(a) Part of speech tagging
(b) Parsing – constituency and dependency
(c) Predicate-argument structure (semantic role labeling)
(d) Named entity recognition
(e) Word sense disambiguation
(f) Relation discovery and classification
(g) Discourse parsing (text cohesiveness)
(h) Language generation
(i) Machine translation
(i) Summarization
(ii) Creating datasets to be used for learning, a.k.a. computable gold annotations

(iii) Active learning
(2) Methods
(a) Principled approaches
(i) Linguistic theory
(ii) Computational science
(b) Machine Learning
(i) Supervised
(ii) Unsupervised
(iii) Lightly supervised
(c) Rules derived by domain experts
(d) Combination
(e) How to integrate knowledge-based information with data-driven methods

iii) Available gold annotations: clinical narrative
(1) MiPACQ (Multi-source Platform for Answering Clinical Questions)
(a) 120K words of clinical narrative
(b) Layers of annotations – pos tags, treebanking, propbanking, UMLS entities and modifiers, UMLS relations and modifiers, coreference
(2) ShARe (Shared Annotated Resources)
(a) 500K words of clinical narrative
(b) Layers of annotations – pos tags, phrasal chunks, UMLS entity mentions of type Disease/Disorder and modifiers
(3) i2b2 (Informatics for Integrating Biology and the Bedside) shared tasks
(a) Medication
(b) Coreference
(4) SHARPn
(a) 500K words of clinical narrative
(b) Layers of annotations – pos tags, treebanking, propbanking, UMLS entities (Diseases/disorders, Signs/Symptoms, Procedures, Anatomical sites, Medications) and modifiers, UMLS relations (locationOf, degreeOf, resultsOf, treats/manages) and modifiers, coreference, template (Clinical Element Model; http://intermountainhealthcare.org/cem)
(5) THYME (Temporal Histories of Your Medical Events)
(a) 500K words of clinical narrative
(b) Layers of annotations – same as MiPACQ and SHARPn + temporal relations (ISO TimeML extensions to the clinical domain)

iv) *Milestones Reached*:
(1) cTAKES Software enhancements
(a) Integration with ClearTK (NLP Machine Learning (ML) package from University of Colorado)
(b) NLP modules trained on clinical data for semantic processing of the clinical narrative (dependency parsing, semantic role labeling, assertion extraction, asserting the subject of a medical condition)
(c) V2.5 released in April, 2012
(d) V2.6 to be released in July, 2012

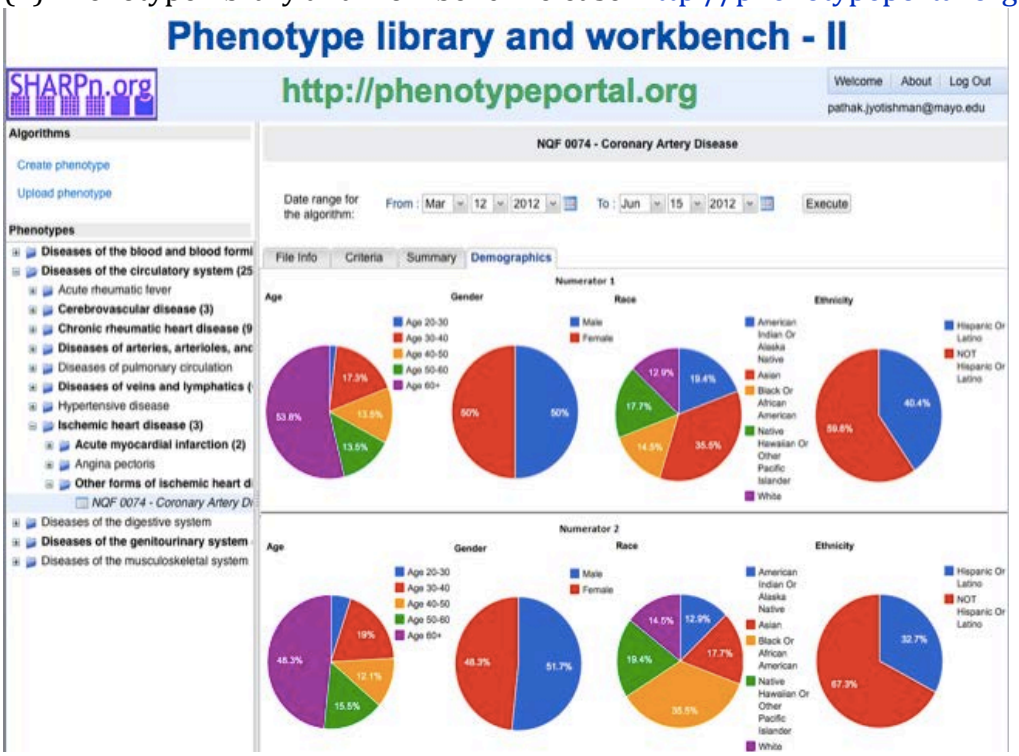
(e) Negotiated with the National Library of Medicine the bundling of UMLS dictionaries with cTAKES releases to facilitate easy deployment and adoption
(f) Graphical User Interface (GUI) for cTAKES to facilitate deployment by non-developers and non-NLPers (https://ohnlp.svn.sourceforge.net/svnroot/ohnlp/branches/cTAKES-GUI-0.0.1/ctakes-gui-0.0.1.zip)
(g) Migration of cTAKES to the Apache Software foundation as an Incubator project to facilitate national and international adoption and contributions (http://incubator.apache.org/ctakes/)
(h) Alpha version of the NLP evaluation workbench

v) Next steps:
(1) Sustained quarterly cTAKES releases with refined NLP modules
(2) Serialization of the populated CEM-based template
(3) Incorporation of NLP extracted data with other data streams for use case consumption, e.g. phenotyping
(4) Methods refinement for semantic processing of the clinical narrative based on gold standard data
(a) Named entity recognition and Word Sense Disambiguation
(b) Relation extraction for finding severity, body site, associated signs/symptoms, indications and treatments
(c) Discovering uniquely identifiable patient events through coreference
(d) Discovering the subject/experiencer and other "roles" of a clinical condition
(e) Portable general purpose sectionizer
(f) Discovering associated negation and uncertainty of clinical mentions
(g) Discovering the full signature of medication mentions

a) High throughput Phenotyping (HTP)
i) *Overarching Goal*: To develop high-throughput automated techniques and algorithms that operate on normalized EHR data to identify cohorts of potentially eligible subjects on the basis of disease, symptoms, or related findings
ii) Current Project Themes
(1) Standardization of phenotype definitions
(2) Library of phenotyping algorithms
(3) Phenotyping workbench
(4) Machine learning techniques for phenotyping
(5) Just-in-time phenotyping
iii) Accomplishments
(1) Automatic translation from NQF QDM criteria to Drools
(a) "executable" Drools flow
(2) Used Clinical Element Models to provide a structure that can be used to encode standard terminologies

(3) Leveraged the NQF Data Model – that allows leveraging meaningful use data (Phase 1 quality measures) vs. data definitions
   (a) Includes a measure authoring tool that allows view/comparisons across different data sets
   (b) Generating rules to implement on top of the data models
   (c) Investigated the JBoss management system which has been effective in healthcare and financial industries
   (d) Developed National Library – that allows data query
   (e) Leveraged the SHARP cloud – secure VPN to execute and to avoid hacking
   (f) Research related activities – experimenting with machine learning and associated rule mining - to determine identification of phenotype definition criteria and work flow presentation including algorithms for decision making
   (g) Active collaboration with Mayo's Transfusion Medicine Group – this involves going back into the clinical data and running queries. This is now being done proactively, and includes a decision support based system to actively monitor patients after blood transfusions to determine adverse events and to support active surveillance.
   (h)

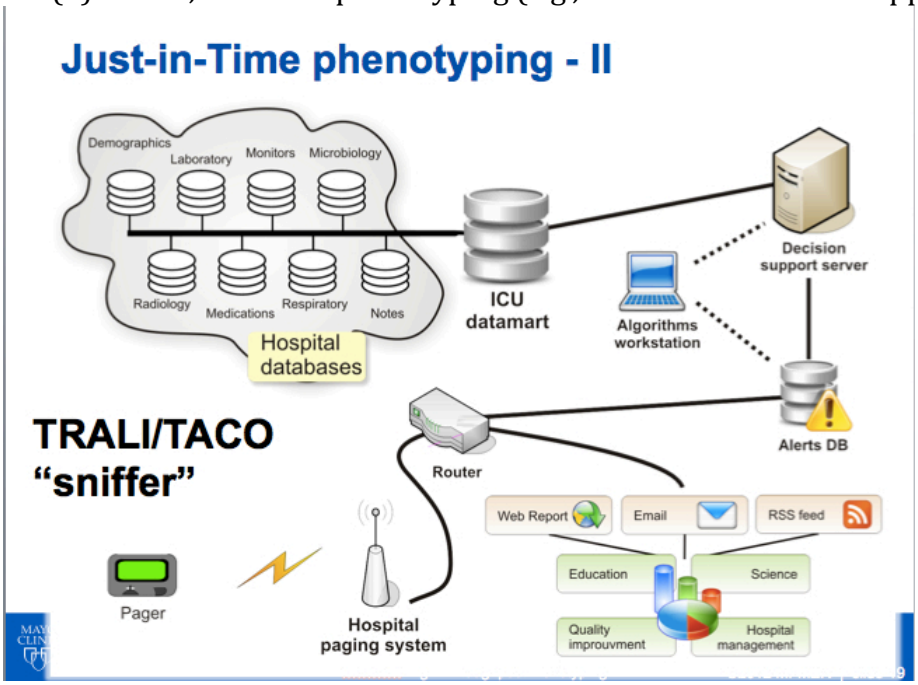
(4) Phenotype library and workbench release: http://phenotypeportal.org



   (a) Converts QDM to Drools
   (b) Rule execution by querying the CEM database
   (c) Generate summary reports

(5) Machine learning and association rule mining
   (a) leverage machine learning methods for rule/algorithm development, and validate against expert developed ones

(6) Just-in-time Phenotyping
(a) Apply algorithms as "data sniffers" that can be plugged within an UIMA pipeline
(b) Online, real-time phenotyping (e.g., for clinical decision support)



iv) 2012 Milestones
(1) Machine learning techniques for algorithm definitions
(2) Online, real-time phenotype execution
(3) Collaboration with NQF, Query Health and i2b2 infrastructures
(4) Use cases and demonstrations
(a) MU quality metrics (w/ NQF, Query Health)
(b) Cohort identification (w/ eMERGE, PGRN)
(c) Value analysis (w/ Mayo CSHCD, REP)
(d) Clinical trial alerting (w/ Mayo Cancer Ctr./CTSA)

b) Infrastructure & Scalability

i) *SHARPn Cloud Resource*
(1) The cloud environment exists to fulfill the mission of SHARPn
(a) Infrastructure as a Service (IAAS)
(b) Ubuntu Enterprise Cloud (UEC) is Ubuntu's Eucalyptus-powered cloud
(2) Cloud Computing Benefits
(a) Remove the cost of acquisition, install, and configuration
(b) Flexibility of Volumes (for example)
(i) Can move to bigger instance
(ii) Backs up only the pertinent data
(3) Setting Expectations
(a) Our cloud is a protected, non-public resource

(i) Government regulations
(ii) Mayo Clinic security policies
(b) Our cloud is a research system not a production system
(i) The cloud is not monitored 24x7
(ii) Maintenance is completed during business hours
(iii) Users setting up instances must have Linux know-how and some system administration skills
(iv) Data backup is available but must be set up by the user
(4) http://informatics.mayo.edu/cirruswiki/index.php/Cloud_Resource_Lab

ii) *SHARPn Pipeline Accomplishments*:
(1) Implement in UIMA (Unstructured Information Management Architecture)
(2) Architecture revision to make pipeline highly configurable
(a) Enable seamlessly pipeline integration of components from different data sources (e.g., NLP)
(b) data sources – HL7, CCD, CDA, and Table format
(c) Model mappings (different EMR systems may have different formats)
(d) Terminology mappings
(e) Inference mappings – infer ingredients from clinical drugs
(3) Generic pipeline components
(a) XML Initializer
(4) Iterative data runs were conducted to improve the pipeline
(a) Various sources were utilized to obtain Medication information.
(b) Legacy application still leverage HL7 2.x messages to convey order information between systems.
(c) CDA narrative and structured documents are coming on-line which convey snapshots of patient's current state and medications.

iii) *SHARPn Pipeline Lessons Learned*
(1) The design of the pipeline needs to be flexible enough to accommodate all kinds of changes – (agile)
(2) UIMA is a nice architecture
(a) Configurable
(b) Model-driven
(i) e.g., taking an XSD specification of CEM and translating into UIMA types
(c) Seamless integration with NLP pipeline
(3) Diverse input formats
(a) Structured -- semantics may be different from different institutions
(i) Needs to understand the data
(b) Unstructured - there is a gap between the semantics of free text and the semantics of standards
(i)
(4) Too many standards to choose when implementing HL7 standards
(a) Mapping from local codes to standard value sets – non-trivial
(5) Versioning of standards is crucial
(a) Do not assume the mapping will be trivial if the EMR data has already adopted the same standard as SHARPN value sets

(6) Different granularities between CEMs and original structures
(a) Dose Strength "50-mg"
(b) NotedDrug CEM: Unit=MG Value=50
(7) Inference
(a) TakenDoseUpperLimit needs to be inferred from TakenDoseLowerLimit
(8) Relational structure for the Demographics data worked well
(a) This provided a nice view into the patients data without having to have a lot of knowledge of the Patient CEM structure.
(b) Allowed for both adds and updates
(c) Was not intended to be a full Master Patient Index (MPI) but does meet the minimal need of linking a given patients records for a given institution.
(9) XML Sample data for the Clinical CEMs
(a) XML samples proved very valuable for validation against the XSD's and for providing an initial set of test messages to Channels.
(b) The more complete these records were the more useful.
(c) Assumed all mappings to code sets were done prior to receiving on the CEM to DB channel.
(10) Code re-use across the Clinical CEM Channels proved very useful
(a) Standardized CEM DB Structure - IndexData, SourceData, and PatientData
(b) Common patient "matching" code used
(c) Currently only supports Add but Updates now possible due to SourceSystemID that was added recently
(11) Mirth support for XML, HL7, etc proved very useful for traversal of structures in code and for field validations.
(12) Supporting Add and Updated for Patient (base) record was useful.
(13) Always create the Patient base record first regardless if Patient or Clinical CEM was received as first CEM to the DB.
(14) Issues and Challenges
(a) Date formatting was one example of needing to understand how the data was being received and used.
(b) Field level storage VS storing of full XML - Tradeoff - Decided to always store full XML – May need to look at additional relational fields on clinical CEMs for better searching support

iv) *SHARPn Pipeline Future Work*:
(1) Integrate NLP normalization pipeline into data normalization
(2) GUIs to simplify the configurations of the pipeline
(a) syntactic mapping
(b) Semantic mapping
(3) Integrate the two options provided to process data into a single project

v) CEM to DB – Accomplishments
(1) Completed the following CEM to DB Mirth Channel Development
(a) Patient CEM Channel (Add and Update)
(b) NotedDrug CEM Channel (Add)
(c) AdministrativeDiagnosis CEM Channel (Add)
(d) Lab (Quantitative, Narrative, Ordinal, and Titer) CEM Channel (Add)

(2) Completed CEM DB Design and Development
(a) Patient CEM Tables
(i) Demographics including Names, Addresses, Telecom, and Language
(ii) Patient Cross Reference and External IDs
(b) Clinical CEM Tables
(i) Index Data
(ii) Source Data
(iii) Patient Data

vi) HL7 to XDR Channel - Accomplishments
(1) Completed the following HL7 to XDR channels for data exchange
(a) HL7 AdminDX to XDR Channel (Sender)
(b) HL7 Lab to XDR Channel (Sender)
(c) HL7 Noted Drug to XDR Channel (Sender)
(d) HL7 Message Separator (Sender)
(e) HL7 Receive Message Separator (Receiver)
(f) ReceiveXDR and Extract HL7 (Receiver)

vii) Next Steps for Channels and CEM DB
(1) Complete the Update feature for Clinical CEM Channels
(2) Enhance the Error Handling and resending of error Message for channels
(3) Additional Relational field data storage for Clinical CEM Messages
(4) Support CEM changes and new data types

viii) Pan-SHARP Collaboration
(1) PanSHARP organization: Distributed team, highly collaborative, and are leveraging existing SHARP technology as much as possible
(2) SHARPn provided Data models and normalized data for the Pan-SHARP medication reconciliation alpha phase.
(a) Identifed a cohort of REP patients meeting visit criteria (Olmsted County residents, received care at Mayo and OMC, history of visits at both institutions in study period (2008-2010), history of multiple visits to Mayo.
(b) Randomly sampled 400 patients from this pool for study purposes.
(c) Obtained medical relevant data sources from both organizations
(i) Demographics (age, date of visits)
(ii) Gender
(iii)Medical History in both organizations.
(d) Merged patients using REP identifiers, then assign sequential study patient ID: 001-400 (begin the anonymization)
(e) Used SHARPn resources to normalize all data sources into a common format (CEMs – clinical element models), which also implies vocabularies like RxNorm for drug names.
(f) The only demographic retained was sex and age. The former was truncated to closest decade, and then synthesized with a random offset between +/- 5 years.
(g) Date information is important for reconciliation. We randomly shifted all dates within patient by +/- 365 days, and then fuzz each individual date by +/- 15 days (that fuzz would have functional impact on "true" reconciliation, but that error is acceptable for this demonstration)

(h) Mayo's SHARPn grant normalized this data into canonical form, where the Texas and Harvard teams then applied a SmartAccess (http://www.smartplatforms.org/) application to execute drug reconciliation.

c) Data Quality

i) Project 1: Data Heterogeneity Study
- (1) Purpose: Compare EHR data between institutions in terms of characteristics (not "quality")
- (2) Institutions: Mayo and Intermountain
- (3) Methods: extract data relative to Type 2 Diabetes from EHR at each institution: diagnoses, labs, meds
- (4) Analysis:
  - (a) Descriptive (compare frequencies and distributions)
  - (b) Tweak selection parameters, and study effects
  - (c) Study within-institution heterogeneity / bias
  - (d) Study differences in institutional source datasets
- (5) Current Status/Milestones
  - (a) IRB, data sharing approval
  - (b) Initial DM2 datasets exist at each institution
  - (c) De-identification (homemade)
  - (d) Initial exchange of de-identified data!
  - (e) Analysis proceeding
    - (i) Comparative analysis of ICD9 codes
    - (ii) Comparison of datasources, missingness
- (6) Next Steps/ Future directions
  - (a) Compare and contrast Mayo and Intermountain data
  - (b) Compare and elucidate idiosyncracies of data sources
  - (c) Draw generalization on heterogeneities
  - (d) Assess impact of these heterogeneities on secondary use
  - (e) White paper

ii) Project 2: Difficult data elements (Body Mass Index, Smoking)
- (1) Purpose: characterize quality aspects of difficult data elements (BMI, smoking…) and develop mitigation or warnings
- (2) Method:
  - (a) extract data (height/weight/BMI, smoking) within Data Heterogeneity study at Intermountain and Mayo
  - (b) Detect errors/ missingness
  - (c) Develop mitigations if possible
- (3) Current status:
  - (a) data related to BMI have been shared, are being analyzed
- (4) Next steps/future directions
  - (a) Comparative analysis of BMI data, data quality / absence issues
  - (b) Smoking status derived from cTAKES-based algorithm about to be reviewed by chart review
  - (c) Develop widgets? White paper

iii) Project 3: Comparison of Computer Algorithm vs. Manual Review for Treatment Cohort selection (a.k.a. the "John Henry" study)
  (1) Purpose: Demonstrate and quantify the cost benefit associated with developing and implementing a ***computer algorithm*** to derive a cohort with high risk type 2 Diabetes compared to ***manual review*** to derive such a cohort. Analyze discordancies between the 2 methods.
  (2) Method: After phased preliminary comparative studies of 20 and 50 potential cases, with refining of algorithm, analyze 200 cases by the 2 approaches. Analyze the discordancies, but also the cumulative costs associated with both methods. Extrapolate to other target sample sizes
  (3) Current Status/Milestones
    (a) First phases complete
    (b) Final contest (200 charts) imminent
  (4) Next steps/Future directions
    (a) Analyze costs using various assumptions
    (b) Report results
    (c) Generalize to other settings?

iv) Project 4: Measuring information in quantitative data
  (1) Purpose: to develop methods to quantify the signal to noise ratio in quantitative data that can be used to inform choices or weights applied to different potential variables related to the same underlying phenotype
  (2) Methods: application of ANOVA to estimate between-subject and within subject components of variance and other methods for estimating signal and noise components, example random capillary glucose and HbA1c
  (3) Current status: gleam in the eye, preliminary proof of concept

2) Program Outputs

a) Products

i) Data Norm - CEM 'Core Models'
http://informatics.mayo.edu/sharp/index.php/CEMS

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 - multiple cTAKES updated releases
http://sourceforge.net/projects/ohnlp/files/icTAKES/

iv) Phenotype library and workbench release: http://phenotypeportal.org

b) 2012 Publications and Presentations

i) PUBLICATIONS

1. Jiang G, Solbrig HR, Chute CG. Using semantic web technology to support ICD-11 textual definitions authoring. ACM International Conference Proceeding Series. 2012; 38-44.

2. Pathak J, Kiefer RC, Chute CG. Applying Linked Data principles to represent patient's electronic health records at Mayo Clinic: A case report. IHI'12 - Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. 2012:455-64.

3. Pathak J, Kiefer RC, Chute CG. The Linked Clinical Data project: Applying Semantic Web technologies for clinical and translational research using electronic medical records. ACM International Conference Proceeding Series. 2012; 94-5.

4. Pathak J, Weiss LC, Durski MJ, Zhu Q, Freimuth RR, Chute CG. Integrating va's ndf-rt drug terminology with pharmgkb: preliminary results. Pac Symp Biocomput. 2012; 400-9. PMID:22174295.

6. Tao C, Wongsuphasawat K, Clark K, Plaisant C, Shneiderman B, **Chute CG**. Towards event sequence representation, reasoning and visualization for EHR data. IHI'12 - Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. 2012:801-5.

7. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, Tao C, Parker CG, Haug PJ, Huff SM, Chute CG. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. J Biomed Inform. 2012 Feb 04. [Epub ahead of print] PMID:22326800. DOI:10.1016/j.jbi.2012.01.009.

8. Pathak J, Kiefer RC, Chute CG. Using Semantic Web Technologies for Cohort Identification from Electronic Health Records to Conduct Genomic Studies. AMIA Summits Transl Sci Proc. 2012; 2012:10-9. Epub 2012 Mar 19. PMID: 22779040. PMCID: 3392057.

9. Sohn S, Wu ST, Chute CG. Dependency Parser-based Negation Detection in Clinical Narratives. AMIA Summit on Clinical Research Informatics (CRI). San Francisco, CA. 2012 Mar.

10. Song D, **Chute CG**, Tao C. Semantator: Annotating Clinical Narratives with Semantic Web Ontologies AMIA Clinical Research Informatics. Mar 2012.

11. Wu ST, Liu H, Li D, Tao C, Musen MA, Chute CG, Shah NH. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus

analysis. J Am Med Inform Assoc. 2012 Jun 1; 19(e1):e149-56. Epub 2012 Apr 04. PMID:22493050. DOI:10.1136/amiajnl-2011-000744.

12. Sohn S, Torii M, Li D, Wagholikar K, Wu S, Liu H. A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes Biomedical Informatics Insights. 2012(5 Suppl 1):43-50.

13. Wagholikar K, Torii M, Jonnalagadda S, Liu H. Feasibility of pooling annotated corpora for clinical concept extraction. AMIA Summit on Clinical Research Informatics (CRI). San Francisco, CA. 2012.

14. Wagholikar KB, Maclaughlin KL, Henry MR, Greenes RA, Hankey RA, Liu H, Chaudhry R. Clinical decision support with automated text processing for cervical cancer screening. J Am Med Inform Assoc. 2012 Apr 29. [Epub ahead of print] PMID:22542812. DOI:10.1136/amiajnl-2012-000820.

15. Jonnalagadda SR, Li D, Sohn S, Wu ST, Wagholikar K, Torii M, Liu H. Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules. J Am Med Inform Assoc. 2012 Jun 16. [Epub ahead of print] PMID:22707745. DOI:10.1136/amiajnl-2011-000766.

16. Liu H, Wagholikar K, Wu S. Using SNOMED CT to encode summary level data - a corpus analysis. AMIA Summit on Clinical Research Informatics (CRI). San Francisco, CA. 2012 Mar.

ii) **FORTHCOMING at AMIA Symposium in November**

1. Liu H, Wu ST, Li D, Jonnalagadda S, Sohn S, Wagholikar K, Haug PJ, Huff SM., Chute CG, Towards a semantic lexicon for clinical natural language processing Accepted paper 2012 AMIA Symposium Nov 2012.

2. Liu M, Shah A, Jiang M, Peterson N, Dai Q, Aldrich M, Chen Q, Bowton E, Liu H, Denny J, Xu H A Study of Transportability of an Existing Smoking Status Detection Module across Institutions Accepted paper 2012 AMIA Symposium Nov 2012.

3. Thompson WK, Rasmussen LV, Pacheco JA, Peissig PL, Denny JC, Kho AN, Miller A, Pathak J. An Evaluation of the NQF Quality Data Model for Representing Electronic Health Record Driven Phenotyping Algorithms. Accepted paper 2012 AMIA Symposium Nov 2012.

4. Pathak J, Kiefer RC, Bielinski SJ, Chute CG. Mining the Human Phenome using Semantic Web Technologies: A case study for type 2 diabetes. Accepted paper 2012 AMIA Symposium. Nov 2012.

5. Li D, Shrestha G, Murthy S, Sottara D, Huff SM, Chue CG, Pathak J. Modeling

and Executing Electronic Health Records Driven Phenotyping Algorithms using the NQF Quality Data Model and JBOSS Drools Engine. Accepted paper 2012 AMIA Symposium Nov 2012.

iii) **Abstracts - FORTHCOMING**

1. Li DC, Shrestha G, Murthy S, Sottara, Huff SM, Chute CG, Pathak J. Applying JBoss® Drools Business Rules Management System for Electronic Health Records Driven Phenotyping. AMIA Annual Symposium 2012. Nov 2012. [Abstract]
2. Pathak J, Al-Kali A, Talwalkar JA, Kho AN, Denny JC, Murphy SP, Bruce KT, Durski MJ, Chute CG. Using Electronic Health Records to Identify Patient Cohorts for Drug-Induced Thrombocytopenia Neutropenia and Liver Injury. AMIA Annual Symposium. 2012 Nov.. [Abstract]
3. Pathak J, Kiefer RC, Freimuth RR, Bielinski SJ, Chute CG. Mining Genotype-Phenotype Associations from Electronic Health Records and Biorepositories using Semantic Web Technologies (poster). AMIA Annual Symposium.. Nov 2012. [Abstract]
4. Shrestha G, Murthy S, Li DC, Hart LA, Chute CG, Pathak J. Visualization and Reporting of Results for Electronic Health Records Driven Phenotyping using the Open-Source popHealth Platform (poster). AMIA Annual Symposium.. Nov 2012. [Abstract]

iv) **Presentations**

1. Podium Presentation Applying Linked Data Principles to Represent Patient's Electronic Health Records at Mayo Clinic: A Case Report American Medical Informatics Association (AMIA) Clinical Research Informatics Symposium 03/2012
2. Podium Presentation Integrating VA's NDF-RT Drug Terminology with PharmGKB: Preliminary Results Pacific Symposium on Biocomputing 01/2012
3. Podium Presentation Using Semantic Web Technologies for Cohort Identification from Electronic Health Records to Conduct Genomic Studies ACM International Health Informatics Symposium 01/2012

3) 2012 Events

2/20/2012-2/24/2012: Healthcare Information and Management Systems Society (HIMSS); Las Vegas, NV
SHARPn tools in the Exhibition with other SHARP and ONC programs.

4/11/2012 – 4/13/2012: ONC Standards and Interoperability (S&I) Framework Face-to-Face (F2F) Conference, Alexandria, VA
SHARPn members in attendance participating in the working sessions.

6/11/2012 – 6/12/2012: Area 4 SHARP Face to Face Conference; Rochester MN
125 Attendees with representatives from SHARPn, other SHARP programs, ONC, FSC, PAC, and Beacons.

4) Partnerships / Relationships / Alliances Maintained

a) 360Fresh
b) 3M Health Information Systems
c) Clinical Information Modeling Initiative (CIMI). International consensus group with Detailed Clinical Models.
d) Consortium for Healthcare Informatics Research (CHIR)
e) Informatics for Integrating Biology and the Bedside (i2b2)
f) Electronic Medical Records and Genomics (eMERGE) Network
g) Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC)
h) Multi-source Integrated Platform for Answering Clinical Questions (MiPACQ)
i) National Library of Medicine, UMLS team
j) National Quality Forum / MAT User Group
k) Open Health Natural Language Processing (OHNLP)
l) Pharmacogenomics Research Network (PGRN) – PGPop team
m) popHealth
n) Southeast Minnesota Beacon Community
o) Standards & Interoperability (S&I) Framework - Query Health
p) Substitutable Medical Apps, reusable technologies (SMArt); Harvard SHARP Team

5) Operational Activities

a) Face-to-face collaboration has been fostered in both intra-SHARP and cross-SHARP program's in cross-knowledge pollination and collaboration activities.
b) 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.
c) IRBs have been submitted and approved at all applicable sites.
d) Data Sharing issues have been resolved with best practice sharing and inventory of existing agreements between institutions reviewed.

6) Personnel / Hiring (ARRA Report)

Budgeted Personnel have remained consistent with justification approved. No significant changes forecasted for Q3/4 of 2012

Calendar Year / Quarter: 2012/ 1
Number of Jobs Count: 19.06
Description of Jobs Created: Imaging Analyst, Project Mgrs., Analyst/Programmers, Investigators,Technologist, Residents, HIE/Terminology Solution ArchitectChief

Executive Officer, Chief Operating OfficerPrincipal Investigator, Data Manager I, Physician, Lead Application Specialist, Research Fellows, Investigator, Project Manager, Programmer Analyst, Research Specialists, Research Staff MemberMedical Vocab Engineer-Stf, Medical Informaticist-Cnslt, Medical Informaticist Sr., Chief Medical Info Officer, Medical Informaticist Sr, Data Architech-Sr, Medical Knowledge Engr-Stf, Project Management-Staff, Medical Vocab Engineer-SrPost-Doctorate Researcher, Fellow, Principal InvestigatorProject Support Specialist, Research Aide, Reimbursement to SUNY for IFR, Investigators, Post Doctorate Researchers, Research Assistant/Associates, NonePost-Doctorate, Principal Investigator, Technical Professional, Senior Manager, Manager, Senior Consultant

Calendar Year / Quarter: 2012 / 2
Number of Jobs Count: 19.06
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

7) Grants Management

Expenditures have remained consistent with work scope approved.
No significant changes forecasted for Q3/4 of 2012.

	Calendar Year / Qtr	
	2012/1	**2012/2**
Direct	$ 4,873,917.00	$ 5,945,731.00
Indirect	$ 1,085,262.00	$ 1,449,226.00
Total Expenditure from Inception	$ 5,959,179.00	$ 7,394,957.00