Team Project: Next Generation Health Exchange

Sharon C. Perelman

Northwestern University School of Continuing Studies

HIT Technology Integration, Interoperability, & Standards

MED INF 405

Guilherme Del Fiol, M.D., Ph.D.

March 4, 2012

Team Project: Next Generation Health Exchange

Introduction

Medical terminology is increasing exponentially. A monolithic terminology standard does not exist and may be impossible to create. Further, terminology users may have different needs for the information, such as administrative versus clinical. Consequently, there are numerous disparate medical terminologies and messaging systems that are often not well suited for health information exchange. The solution is to develop a system that normalizes the disparate nature of medical language and messages for health information exchange (HIE). The Mayo Clinic is developing the next generation HIE system. Our intent is to evaluate the program in terms of benefits, challenges, risks, architecture, data exchange, information flow, stakeholder impacts, and current state. We also intend to provide recommendations.

System Description/ Benefits

Traditionally, a patient's medical information, such as medical history, exam data, hospital visits, lab reports, radiology results, and physician notes, are inconsistently recorded and stored in multiple locations electronically and non-electronically. Healthcare data is created in multiple formats, such as unstructured narrative text, structured and coded data, images, sounds, waves, etc. Most patient care data is captured in unstructured narrative text that consequently locks information in free text. There may also be usability issues associated with clinicians not being able to find codes for every single clinical feature of their patient's condition. Medical language is often ambiguous and imprecise. The number of medical concepts and terms are an order of magnitude of 6 digits and are growing exponentially with genetics. Additionally, there are too many standards to choose from and not many widely adopted. This heterogeneity inhibits secondary use. If terminologies could be standardized or normalized the benefits of secondary use would enable:

- Clinical decision support
- Data analysis for quality assurance, financial planning, auditing, etc.
- Public health reporting
- Clinical research
- Multiple transactions, such as order entry, ADT, results reporting, etc.

Although inconvenient for the computer, natural language is powerful for humans. It captures nuances that even the best standard terminology will never capture. In December of 2010, the Office of the National Coordinator (ONC) announced the Strategic Health IT Advanced Research Projects (SHARPn) as part of the federal stimulus project. One of the four-awarded projects (Area 4) focuses on EMR secondary data use. As previously stated, secondary use data enhances patient safety and enables clinical quality metrics, development programs, clinical decision support and practice variation monitoring. The clinical and translational research are also dependent on effective secondary use of clinical information, including clinical trials, observational cohorts, outcomes research, comparative effectiveness, and best evidence discovery. Area 4, headed by The Mayo Clinic, consists of a collaboration of 16 academic and industry partners tasked to develop tools and resources that influence and extend secondary uses of clinical data. The program consists of six projects that are strongly interrelated and mutually dependent. This paper will focus on three of the six areas: specifically: (1) Semantic and Syntactic Data Normalization, (2) Natural Language Processing (NLP), and (3) Phenotyping Applications (Chute et al., 2011).

Semantic and Syntactic Data Normalization

Data Normalization

Data normalization converts similar data expressed in ***different*** formats to a single format. For instance, age is usually expressed in years for adults, but in months or days for toddlers and babies. Data normalization

would convert all age data to a single scale, be it in days or months. This provides a standard structure and terminology necessary for clinical decision support (SHARPn.org, 2011). Clinical data normalization provides the potential to:

- build a data normalization pipeline
- establish a global resource for health technology and value sets
- establish a modular library for normalization algorithms
- test, evaluate and revise normalization pipelines
- identify algorithms in EMR and normalize them against Clinical Element Model Systems (CEMS).

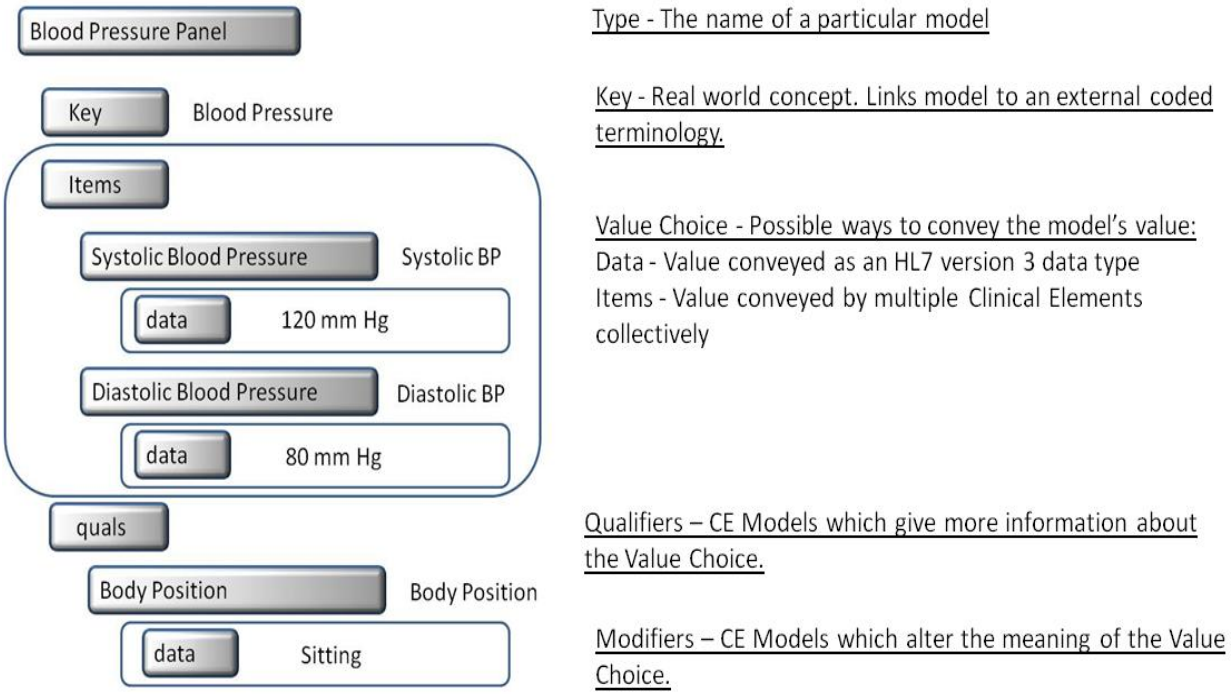
In order for secondary use data to be practical, applications need to be efficient and reliable and data must be comparable and consistent**.** The SHARPn program suggests that heterogeneous data captured from disparate EHR systems may be **comparable and consistent** with post-hoc normalization. SHARPn uses the Unstructured Information Management Architecture (UIMA) framework that facilitates post-hoc normalization. The UIMA framework is an open scalable and extensible platform for building analytic applications or search solutions that process text or other unstructured information to find the latent meaning, relationships and relevant facts buried within. UIMA enables developers to build analytic modules and to compose analytic applications from multiple analytic providers, encouraging collaboration and facilitating value extraction for unstructured information. Normalization requires that two dimensions be satisfied, consisting of syntactic and semantic elements. Semantic generally refers to meaning and Syntactic refers to arrangement or structure. The SHARPn project relies heavily on data normalization foundational worked performed at the Regenstrief Institute and the Indiana Health Information Exchange (HIE). Regenstrief has developed a normalization pipeline called Health Open Source Software (HOSS). HOSS excels at making ill-formed HL7 message contents into well-formed structures. SHARPn's added value is to modularize and simplify normalization task elements to develop a high-throughput UIMA pipeline. Mayo uses its own open-source (LexGrid) to enhance semantic normalization. Mayo believes that additional specificity is needed to address the spectrum of use-cases in secondary data use and the clinical granularity of information being generated in today's EHRs. Semantic normalization depends upon the creation or availability of mapping files, for example local laboratory codes into LOINC. Syntactic and semantic normalization requires that a common form (canonical) representation be specified as a target for normalization activities.

Canonical representation can best be depicted in detailed clinical models (CMs) because CMs retain computable meaning when data is exchanged from disparate systems. Huff suggests that CMs need to be comprehensive, flexible and extensible. CMs must accommodate full patient representation as well as preserve headroom to incorporate additional elements and attributes without requiring underlying changes to software or database. CMs need to comply with existing formal schema such as XML constraints relative to structure and content requirements without modification. CMs need to have tight linkage to standard terminologies such as SNOMED CT, LOINC, HL7, etc. CMs must posses a mechanism for determining when something is not observed or not present. Clinical model version control is also important to determine which model is in effect when the data was stored. CMs must allow values (number of elements of the set) to be easily modified (Huff Stan, 2008).

Huff designed the Clinical Element Models (CEMs) to represent clinical data model. The CEM is a strategy designed to represent logical models for clinical data elements to ensure unambiguous data representation, interpretation, and exchange within and across heterogeneous sources and applications. CEMs are the combination of an abstract instance model and an abstract constraint model. The abstract instance model defines a structure to represent instances of medical data, and the abstract constraint model defines constraints on values in the abstract instance model. Together these models provide context. Context knowledge is needed in order to retain the computable meaning of data when exchanged. For example, knowing the context of a blood pressure reading might be important. Was the patient setting or standing? Or running prior to the reading? Was the cuff placed on the upper arm or the ankle? The diastolic pressure readings may be equivalent between the arm and ankle blood pressure readings but the systolic pressures may not be equivalent. This might reflect the fact that the ankle systolic blood pressure is physiologically

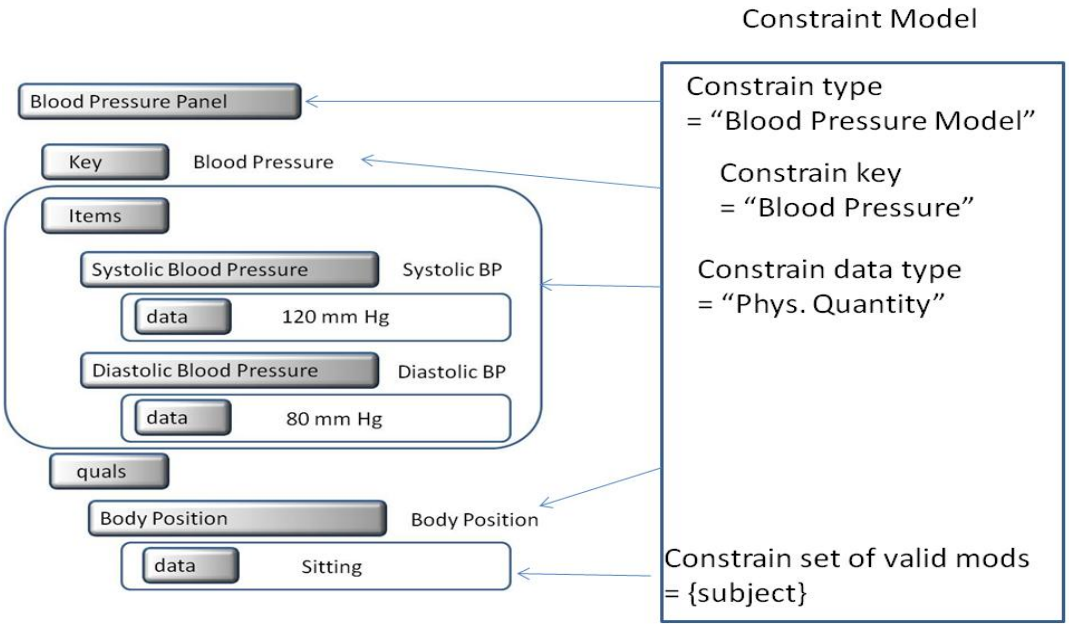
higher than the arm systolic blood pressure. The data could be subject to misinterpretation if context is unknown. The abstract instance model is provided below in Figure 1 using the blood pressure reading example. As shown, the model type is the blood pressure panel and the real word concept or key is blood pressure. The value choice depicts the blood pressure values in systolic and diastolic at 120 mm Hg and 80 mm Hg respectively. The qualifiers provide additional contextual information such as body position and modifiers such as setting.

Figure 1- Abstract Instance Model -Modified
from (Huff, Stan 2008)



The abstract constraint model is depicted below in Figure 2. SHARPn uses CEMs as its canonical representation. Presently, over 4,000 XML schema for CEMs, such as blood pressure measurement or specific laboratory tests, are defined. SHARPn, and a growing consortium of informatics users, are contributing to the CEM library, which is an open-source artifact.

Figure 2- Abstract Constraint Model



Natural Language Processing (NLP)

SHARPn is also using NLP technology. NLP captured wide public attention with IBM's Watson (artificial intelligence) that appeared on the Jeopardy quiz show. Watson answered questions in natural language and consistently performed at the level of human experts in terms of precision, confidence and speed against two of the best-known Jeopardy Champions. Watson had access to 200 million pages of structured and unstructured content.

The benefits of NLP include information exchange, transformation of unstructured text to structured representations, and the ability to merge clinical data that is extracted from the EHR in free text form with structured data. Not only is NLP a necessary tool on the road to phenotyping if the goal is to extract patient information from free text, it also has shown benefits of its own. In one study comparing NLP to patient safety indicators using discharge coding, NLP showed more sensitivity by far in finding post-operative surgical complications than did the coding. NLP correctly identified acute renal failure 82% of the time, compared with 38% using patient safety indicators; for pneumonia, the numbers were 64% for NLP and 5% for patient safety indicators; for sepsis, 89% vs. 34%. Though specificity was slightly lower with NLP, it still remained high (Murff, et al., 2011). In a small 150 patient study NLP was found to be as good as human coders in assessing clinician adherence to tobacco treatment (Hazlehurst, et al., 2005).

NLP has not been used extensively because it has been time-consuming and error-prone. Experts with a great deal of domain and NLP knowledge must preprocess the free text, submit it through a command-line, and interpret outputs with post-processing scripts. As a result, it is expensive to use and few applications currently use NLP (Chard, Russel, Lussier, Mendonca, & Silverstein, 2011) cTAKES, the functionality developed by SHARPn and explained more in detail below, is making NLP much easier to use.

SHARPn uses NLP as an enabling technology for phenotype extraction from clinical free text. NLP is used as a means of Information Extraction (IE) transformation of unstructured free text into structured representations. The clinical data extracted from free text is then merged with structured data. The goal is to research and implement general-purpose modular solutions for the discovery of key components to be used in a wide variety of biomedical use cases. SHARPn's efforts are on methodologies for clinical event discovery and semantic relations between these events. Subsequently, the discovered entities and relations will populate normalization targets in the CEMs where each CEM is summoned through the SNOMED CT or RxNorm code. The process involves the following steps:

- Find the information nuggets, usually in the form of sentences, in the free text
- Find the sentence boundary
- Find the tokens
- Assign part of speech to each token
- Unify tokens into phrases (units such as noun phrases, prepositional phrases, word phrase)
- Noun phrases become the lookup term to search for the relevant clinical event
- Once a clinical event is identified, it is mapped to an ontology such as SNOMED-CT
- Based on the clinical element model, the attribute related to each clinical event is found (SHARPn.org, 2011).

As previously mentioned, the engineering framework within which the NLP project functions is UIMA. Core technologies such as clinical Text Analysis and Knowledge Extraction System (cTAKES) built within UIMA provide a solid basis for expandability and software development. More specifically, cTAKES is an open-source natural language processing system for information extraction from electronic medical record clinical free-text. It processes clinical notes, identifying types of clinical named entities — drugs, diseases/disorders, signs/symptoms, anatomical sites and procedures. Each named entity has attributes for the text span, the ontology mapping code, context (family history of, current, unrelated to patient), and negated/not negated.

Phenotyping Applications

Phenotyping enables researchers to identify categories of patients with particular characteristics (binning), be it for research, clinical trials, clinical decision support rules, or quality metrics. Phenotyping creates order in masses of patient information (SHARPn.org, 2011). Several studies have shown NLP and phenotyping to be effective in extracting notifiable diseases, co-morbid conditions, medications, and identifying adverse events, as well as identifying patients with rheumatoid arthritis (Liao, et al., 2011). Another study showed that it was possible to identify from EMRs patients with DNA-related diseases, making it possible to obtain cohorts for studies of these diseases from EMRs, rather than from biobanks (Ritchie, et al., 2010). 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. The goal is 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. Typical components include:

- billing and diagnosis codes
- procedure codes
- labs
- medications
- phenotype-specific co-variates (e.g., demographics, vitals, smoking status)

The SHARPn team incorporates Drools for phenotyping. Drools is a business rule management system with a forward chaining inference-based rules engine tailored for the Java language. Drools operates in accordance with production rules which use inclusion and exclusion criteria for cohort identification, numerator and denominator criteria for clinical quality metrics, and trigger criteria for clinical decision support.

Stakeholders

In light of the collaborative effort among several organizations involved in the SHARPn project, there are numerous stakeholders, all with different needs and hopes of what to achieve from this endeavor. First, the healthcare providers are envisioning a system that is precise, accurate and most importantly, user friendly. They have concerns regarding the standardization of nomenclature and its accuracy. Physicians envision that NLP will provide voice recognition that automatically converts unstructured dictated text into structured

binning in the EHR. A goal of NLP would be to advance towards immediate feedback from clinical decision support systems (CDSS) as the physician dictates into the EMR.

The vendors would hope to benefit from the advances and milestones reached via the development of the Next Generation Health Exchange System. These vendors want to partner with the developers to make resources available for commercial deployment and support which will ultimately produce a better product.

Patients look forward to a user-friendly system that is easy to navigate. Privacy and security are of utmost importance to the consumer. These users require easy access to their personal health records via a secure portal, and ultimately are looking for voice input. Along the same line are the benefits to the general public as stakeholders. The overall goal is to engage public health services for underserved populations as well as to create, evaluate and refine informatics information to leverage EHR data and improve public health by utilizing the knowledge gained.

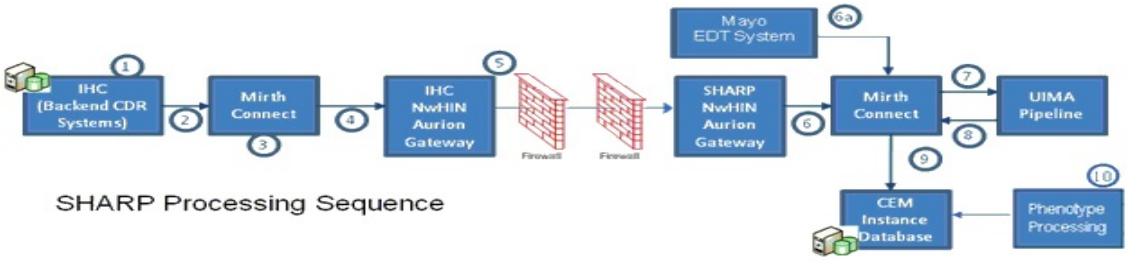
Researchers are anticipating significant advances as a result of the Next Generation Health exchange and NLP advancements. They anticipate the ability to evaluate large quantities of data for quality and insert these findings into the CDSS as well as to use this data to trigger events in the CDSS. NLP will provide a platform for data analysis and assist in evaluating the effectiveness of procedures, providing more accurate outcome evaluations. Researchers aspire to utilize the information from the EHR to aid in clinical trials, but taking it a step further, they desire the ability to extract information and identify possible candidates for these trials.

The final stakeholders are the payers, including Medicare, Medicaid and private payers. Data obtained will allow for further evaluation of clinical outcomes and cost effectiveness of procedures along with evidence-based medicine for best practices.

Information Flow/ Architecture

The architecture is designed to support moving data from providers' clinical databases to UIMA so that it can be parsed, normalized, and stored into CEMs quickly, then output to the CEM Instance Database.

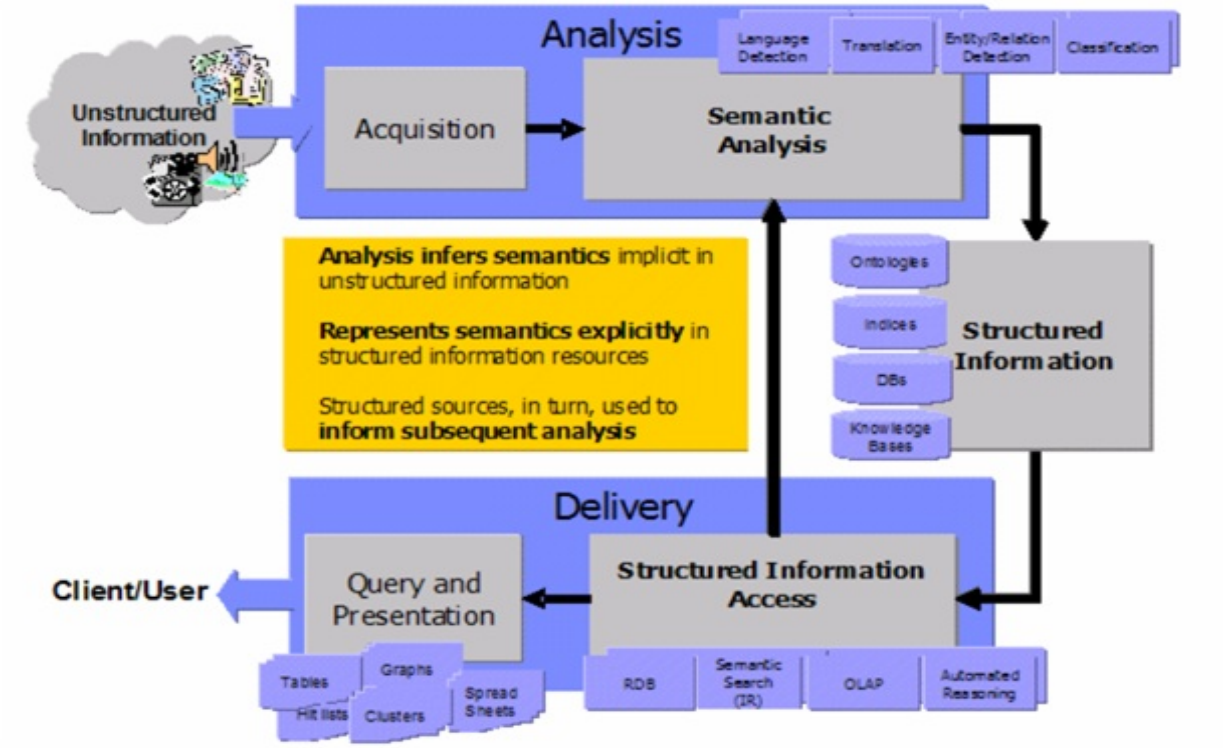
Figure 3 High -Level Architecture
from (SHARPn.org, 2011)



De-identified HL7 v2.x messages, created from the Intermountain Healthcare (IHC) clinical data repository (CDR) (1), are sent to Mirth Connect, a message interface engine (2), which embeds the HL7 messages in a Nationwide Health Information Network (NwHIN) Document Submission (External Data Representation (XDR)) message so that the message can be exchanged between different computer systems (3). The XDR message is sent from Mirth Connect to the NwHIN Aurion Gateway (a health information exchange platform that allows secure message exchange) (4) on the IHC side (5), to the SHARP side, crossing the IHC and Mayo firewalls. The NwHIN Aurion Gateway on the SHARP side picks up the messages and transmits them to SHARPn's Mirth Connect (6). The SHARPn Mirth Connect also receives lab information in HL7 messages, as

well as clinical documents from the Mayo system (6a). These messages are sent to the UIMA Pipeline (7), where they are normalized and transformed into clinical element model (CEM) instances, and sent as XML messages back to Mirth Connect (8), which sends the CEM instances to a database (9). A phenotyping processor works on the stored CEM messages (10) to determine sets of characteristics about each patient.

Figure 4 UIMA Architecture from (IBM)



The purpose of UIMA is to provide bridges between unstructured information and structured knowledge. The Common Analysis Structure (CAS), is used by UIMA to provide read access to the artifact (e.g. document, lab results) being analyzed, and to provide read/write access to the analysis results or annotations associated with the artifact. In the NLP part of the UIMA, the CAS passes from one analysis engine to the next, each performing a different function. For the clinical documents, in the NLP piece of the UIMA, the first analyzer finds sentences; the next finds tokens; the next, a lexical variant generator, attempts to match a token with a term in UMLS, and so on. Once tokens or phrases are identified, other analysis engines within UIMA analyze different facets of the CAS (e.g. medications, smoking status). For each piece of information retrieved, a CEM can be generated, usually in an XML format. Drugs can be mapped to an RxNorm term. Languages used are predominantly C++ and Java (Chute, et al., 2011), (IBM).

Data Exchange

HITECH was enacted as part of the American Recovery and Reinvestment Act (ARRA); its focus is to promote Meaningful Use of health information technology while maintaining provisions to advance the delivery of health care and improve health care outcomes. Another key component of the HITECH ACT is exchanging health care information among providers (SHARPn.org, 2011).

Under the auspices of the Office of the National Coordinator (ONC), The National Information Health Network was established in 2004 as a successor of the National Health Information Infrastructure. The initiative was to

exchange health care information; public concern regarding government access to personal health information led to the name change of this organization to the Nationwide Health Information Network (NwHIN). Viewed as a " network of networks", the NwHIN will tie together HIEs, integrate delivery networks, pharmacies, government agencies, labs, providers, payers and stakeholders. Requirements of a health care organization that will participate in the NwHIN is the acquisition of a unique Organizational Identifier (OID) that will allow the healthcare system or vendor to receive and send messages to trusted entities within the NwHIN through an interface such as CONNECT.

CONNECT is an open source software package that supports HIE at both the local and national level. In 2008 federal agencies implemented a program to connect their health IT systems into NwHIN, as a means of sharing health care data using nationally recognized interoperability standards. This Federal Health Architecture Program now includes at least 33 cooperative members.

A scaled down version of CONNECT called DIRECT allows two entities to share medical records over the Internet. Providers who do not have access to NwHIN 's CONNECT may utilize the system via a secure email address issued by the ONC that would allow sharing of information. DIRECT may facilitate collection of data by government agencies for analysis of health care statistics and evidence-based medicine, along with data sharing between two institutions.

In order to exchange information, providers may use features of the EHR to connect to a Health Information Exchange (HIE); the HIE will then facilitate the exchange with other EHRs or Personal Health Records (PHRs) within that particular HIE or other HIEs via the NwHIN. If the provider does not have an EHR, web access may be utilized to access a portal within the HIE. Consumers may also connect to an HIE which in turn will connect to the NwHIN. The patients may use features of their PHR that they designate as a repository for their personal health information (PHI).

Stakeholders involved in the NwHIN include Care Delivery Organizations (CDO) that use the EHR and consumer organizations that operate PHRs and other consumer applications. The HIEs themselves are multi-stakeholder entities that enable movement of health related data. Organizations may lack the necessary technical and operational infrastructure to conform to the architecture of the NwHIN. In these cases Health Information Service Providers (HSP) will support NwHIN participants by providing them with the organizational and technical infrastructure participants need to access to the NwHIN.

The NwHIN is not an actual physical network that runs on government servers at the Department of Health and Human Services or a large network that stores patient information. The long-term plan is that the ONC will maintain overall responsibility for NwHIN's governance.

In order for information in free text to be exchanged easily, tools such as those provided by SHARPn must be used. Data from disparate sources will have to be parsed and put into bins, then reassembled for later use. Since HIEs and networks of HIEs are in their infancy, the more tools available for sharing data, the better.

Potential Challenges/Critiques

Substantial limitations and challenges persist for SHARPn. Some are listed below:

- "Well-curated semantic mapping tables" are needed between raw data and the target terminologies (Chute, 2011). It is important to have mapping standards so that they are not implemented differently in every application.
- Phenotyping algorithm design is non-trivial, requires significant expert involvement and is a highly iterative process.
- Data access and representation is problematic because of the lack of unified vocabularies, data elements, and value sets. There is also questionable reliability of ICD & CPT codes (miscoding) .
- The technology is still immature and has the status of "middle-ware" and will always limit any highly visible manifestations of the tools to end-users of secondary use applications. This

technology is solving a stopgap or plumbing issue and is consistent with a demonstration program.

- The path toward commercially supportive implementations of SHARPn technologies is still distant from expanding adoption by a broader audience.
- There is a management challenge of integrating the complex organizational and technical structures into a convergent practical and reliable next generation HIE system. The work of the 16 academic and industry partners that are involved in SHARPn must integrate well together.
- Phenotyping algorithms are generally unstructured and have no template; though this makes them amenable to human consumption, they are vulnerable to misinterpretation and open to ambiguity in cohort eligibility. They are also not machine processable. To address this issue, SHARPn is investigating structured eligibility criteria representation models stored in XML, which could be queried and parsed by computer. They would like to create a library of phenotyping algorithms available on the web (Chute, 2011). This is still a labor-intensive solution.
- At this point, NLP and phenotyping are not easily available to larger audiences without significant support. Thus, it will not go into general use until it is easier to use. Perhaps it is best used at this point in HIEs.
- NextGen appears to be geared toward EHRs written just in English. Will the NLP and mapping portion have to be repurposed for other languages? What about different NLP products? Ultimately they should all produce the same CEMs.
- The SHARPn progress report provides a very good listing of notable milestones accomplished and next steps. The report does not provide schedules, however (Chute, 2011). It is difficult to determine if there is a schedule variance or projected schedule variance via performance indicators. Not knowing prevents outsiders from determining program performance.

Recommendations

- Code use methodology needs to be standardized if an institution is to participate in UIMA. The CEMs should have built in accommodations for the variations in amount of use of the codes (ICD9 code use varies by institution, even if the disease prevalence does not).
- Expand the use of codes, and make room for new coding systems, e.g. use of ICD10.
- Set up the mapping tables in a centralized, neutral organization, e.g. the National Library of Medicine, or some group like HL7; also set up centralized, state-run UIMA and databases so that implementation is consistent.
- If the same NLP product is not used universally, thoroughly test all NLP products to make sure that they their end results are equivalent. This should be part of the Evaluation and Framework in SHARPn's Project 6.
- SHARPn needs to bridge the path between demonstration projects to a commercially viable program. Even though the goal is to implement open-source tools, SHARPn needs to validate the commercial viability of these core tools in the market. This can be accomplished in the context of the evaluation framework in Project 6.
- In the next quarterly progress report provide master schedule progress by aggregated tasks, associated percentage complete and spend plan vs. actual. Provide moderate to high project risks to determine risk probability and mitigations.
- Enact rules that clarify privacy issues. Ensure that privacy standards for all data repositories and networks that cross state lines surpass the strictest of the state rulings.

Future State

It is hoped that SHARPn products will enhance clinical decision support, process real-time physician dictation, and provide study cohorts quickly and easily. Public health information mining may be a more widespread earlier use, as data does not have to be 100% accurate when collected in the aggregate. With well-defined and properly populated CEMs, the data-mining possibilities are infinite.

Meaningful Use

The ability to accurately assess quality and performance via data extraction will promote significant advances towards Meaningful Use. NLP will become an effective tool in extracting information from the EHR beyond a standard query. For example, EMRs with NLP processing identified up to 12 times the number of pneumonia cases and twice the rate of kidney failure and sepsis as did searches based on billing codes (Murff, et al, 2011). Thus, uncoded postoperative complications, though they existed, did not show up on quality and performance reports. Ultimately this type of information extraction and utilization will provide global benefits for quality, and cost containment, and ultimately highly improved performance. SHARPn products are well-positioned to enhance Meaningful Use because they provide tools for clinical decision support.

Genomics

DNA sequencing and the rapid advances in genomics are the next frontier in medicine. The implication, however, is that medical terminology will expand exponentially. SHARPn tools have the potential to extract genomic-related information from databases and combine this information with data in the EHR for new and exciting therapies customized to an individual's DNA. UIMA, in its scaleup and limitless scaleout capabilities, is uniquely positioned to handle the enormous volumes of data.

References

Alembic Foundation. (n.d.). *What is Aurion?* Retrieved 02 14, 2012, from Aurion.org: http://aurionproject.org/about/what_is_aurion

Chard, K., Russel, M., Lussier, Y., Mendonca, E., & Silverstein, J. (2011). A Cloud-based Approach to Medical LP. A*MIA Annu Symp Proc. 2011*, (pp. 207-216).

Chute, C. (2011, 12 31). *SHARP 2011 Annual Report.* Retrieved February 27, 2012, from Informatics, Mayo: http://informatics.mayo.edu/sharp/images/5/51/SHARP_Annual_Report_2011_Final.pdf

Chute, C., Pathak, J., Savova, G., Bailey, K., Schor, M., Hart, L., et al. (2011, 10 22). *The SHARPn Project on Secondary Use of Electronic Medical Record Data: Progress, Plans, and Possibilities.* Retrieved 02 14, 2012, from PubMed Central: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243296/

Chute, C., Huff, S. (2010).Strategic Health IT Advanced Research Project Progress Report. Retrieved from www.infomatics.mayo.edu/sharp/images/c/cc/SHARP_PROGRESS_REPORT-2011

Coyle, J., Heras, Y., Oniki, T., & Huff, S. (2008, 11 14). *Clinical Element Model.* Retrieved 02 14, 2012, from Mayo Clinic: http://informatics.mayo.edu/sharp/images/e/e2/CEM_Reference20081114.pdf

Divita, G., Browne, A., & Rindflesch, T. (1998). Evaluating Lexical Variant Generation to Improve Information Retrieval. *Proceedings of the AMIA Synposium*, (pp. 775-779).

Hazlehurst, B., Sittig, D., Stevens, V., Smith, K., Hollis, J., Vogt, T., et al. (2005). Natural Language Processing in the Electronic Medical Record: Assessing Clinician Adherence to Tobacco Treatment Guidelines. *American Journal of Preventive Medicine* , 434-439.

Huff Stan, C. J. (2008, November 14). Clinical Element Model. Retrieved February 17, 2012, from Mayo Informatics: http://informatics.mayo.edu/sharp/images/e/e2/CEM_Reference20081114.pdf

IBM. (n.d.). *UIMA Architecture Highlights.* Retrieved 02 18, 2012, from IBM.com: http://domino.research.ibm.com/comm/research_projects.nsf/pages/uima.architectureHighlights.html

Liao, K., Cai, T., Gainer, V., Goryachev, S., Zeng-Treitler, Q., Raychaudhuri, S., et al. (2011). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res* , 1120-1127.

Mirth Corporation. (2012). *Mirth Commnunity Overview*. Retrieved 02 14, 2012, from Mirth Corporation: http://www.mirthcorp.com/community/overview.

Murff, H., FitzHenry, F., Matheny, M., Gentry, N., Kotter, K., Crimin, K., et al. (2011). Automated Indentification of Postoperative Complications Within an Electronic Medical Record Using Natural Language Processing . *JAMA* , 848-855.

Pathak, J. (2011, June 30). Assistant Professor of Biomedical Informatics. Retrieved February 15, 2012, from SHARP Area 4 – High Throughput Phenotyping: informatics.mayo.edu/sharp/images/0/01/4.SHARP_F2F_Phenotyping

Ritchie, M., Denny, J., Crawford, D., Ramirez, A., Weiner, J., Pulley, J., et al. (2010). Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* , 560-72.

SHARPn.org. (2011, 11 21). *Data Normalization*. Retrieved 02 14, 2012, from SHARPn.org: http://informatics.mayo.edu/sharp/index.php/Data_Normalization

SHARPn.org. (2011, 11 02). *HTP Research*. Retrieved 02 14, 2012, from SHARPn.org: http://informatics.mayo.edu/sharp/index.php/HTP_Research#High-throughput_Phenotyping_.28HTP.29

SHARPn.org. (2011, 06 30). Initial Prototype for Clinical Data Normalization and High Throughput Phenotyping. Rochester, MN, US.

SHARPn.org. (2011, 10 22). *NLP Research*. Retrieved 02 14, 2012, from SHARPn.org: http://informatics.mayo.edu/sharp/index.php/NLP_Research

SHARPn.org. (n.d.). *SHARP CONNECT.* Retrieved 02 14, 2012, from mayo.edu: cs.mayo.edu/sharp/images/3/34/Visio-SHARP_CONNECT_DIAGRAM_(3).pdf

Terry, K. (2011, July 20).Mayo Clinic Builds Next-Gen Health Information Exchange. Information Week Health Care. Retrieved from www.informmationweek.com/news/healthcare/interoperability

The Apache Software Foundation. (2011). *Getting Started: Apache UIMA Asynchronous Scaleout*. Retrieved 02 14, 2012, from Apache UIMA: http://uima.apache.org/doc-uimaas-what.html

The Apache UIMA Team. (2008, 07). *Subject: [ANNOUNCE] UIMA-AS (Asynchronous Scaleout) add-on for UIMA, version 2.2.2-incubating, released - msg#00008*. Retrieved 02 14, 2012, from osdir.com: http://osdir.com/ml/apache.maven.announce/2008-07/msg00008.html

Wikipedia. (2011, 10 05). *External Data Representation*. Retrieved 02 14, 2011, from Wikipedia: http://en.wikipedia.org/wiki/External_Data_Representation