

Fall 11-5-2013

Development and Evaluation of an Ontology-Based Quality Metrics Extraction System

Sina Madani

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthshis_dissertations



Part of the [Bioinformatics Commons](#), and the [Medicine and Health Sciences Commons](#)

Recommended Citation

Madani, Sina, "Development and Evaluation of an Ontology-Based Quality Metrics Extraction System" (2013). *UT SBMI Dissertations (Open Access)*. 28.
https://digitalcommons.library.tmc.edu/uthshis_dissertations/28

This is brought to you for free and open access by the School of Biomedical Informatics at DigitalCommons@TMC. It has been accepted for inclusion in UT SBMI Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.


Development and Evaluation of an Ontology-Based
Quality Metrics Extraction System

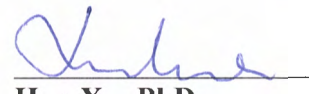
By

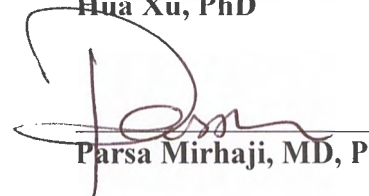
Sina Madani, MD, MS

November 5, 2013

APPROVED:


Dean F. Sittig PhD, Chair


Hua Xu, PhD


Parsa Mirhaji, MD, PhD


Victoria Jordan, PhD


Kim Dunn MD, PhD

Date approved: 11/5/2013

Development and Evaluation of an Ontology-Based
Quality Metrics Extraction System

A
Dissertation

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

By

Sina Madani, MD, MS

University of Texas Health Science Center at Houston

2013

Dissertation Committee:

Dean F. Sittig, PhD¹, Advisor, Chair
Hua Xu, PhD¹, Advisor
Parsa Mirhaji, MD, PhD¹, Advisor
Kim Dunn, MD, PhD¹, Advisor
Victoria Jordan, PhD², Advisor

¹The School of Biomedical Informatics

²MD Anderson Cancer Center, Department of Quality Measurement & Engineering

Copyright by

Sina Madani

2013

Dedication

This thesis is dedicated to my dear parents, who have always been a source of inspiration and encouragement, to my brother, and to my beloved late sister.

I also lovingly dedicate this thesis to my wife, Laila, who supported me unconditionally each step of the way, and to my daughter Alav who brought so much joy and amazement to my life.

Acknowledgements

I would like to thank my committee chair, Dr. Dean F. Sittig who continuously provided assistance and guidance for my thesis and boundless support for my AHRQ fellowship awards. The good advice and encouragement of Dr. Kim Dunn upon my admission to the school of Biomedical Informatics has been invaluable on all academic levels, for which I am forever grateful. I am also thankful to Dr. Hua Xu for being an extraordinary committee member who helped me succeed in my research. In addition, a thank you to Dr. Parsa Mirhaji who introduced me to the Semantic Web technology and whose guidance on “ontologies” had lasting effect. Last but not least, I would like to express my deepest appreciation to my colleague, Dr. Reza Alemy, who continuously conveyed a spirit of adventure in regard to teaching and helping me with computer programming. I thank MD Anderson Cancer Center, Department of Clinical Analytics & Informatics (Dr. John Frenzel) and Department of Quality Measurement & Engineering (Dr. Victoria Jordan) for providing me support and access to the data for my dissertation. I would also like to acknowledge the financial and academic support of the Keck Center of the Gulf Coast Consortia and its staff, particularly for fellowship awards in 2012 & 2013 on the Training Program in Patient Safety & Quality, Agency for Healthcare Research & Quality (AHRQ) T32HS017586, which provided the necessary financial support for this research.

Abstract

The Institute of Medicine reports a growing demand in recent years for quality improvement within the healthcare industry. In response, numerous organizations have been involved in the development and reporting of quality measurement metrics. However, disparate data models from such organizations shift the burden of accurate and reliable metrics *extraction* and *reporting* to healthcare providers. Furthermore, *manual abstraction* of quality metrics and diverse implementation of Electronic Health Record (EHR) systems deepens the complexity of consistent, valid, explicit, and comparable quality measurement reporting within healthcare provider organizations.

The main objective of this research is to evaluate an ontology-based information extraction framework to utilize unstructured clinical text for defining and reporting quality of care metrics that are interpretable and comparable across different healthcare institutions.

All clinical transcribed notes (48,835) from 2,085 patients who had undergone surgery in 2011 at MD Anderson Cancer Center were extracted from their EMR system and pre-processed for identification of section headers. Subsequently, all notes were analyzed by MetaMap v2012 and one XML file was generated per each note. XML outputs were converted into Resource Description Framework (RDF) format. We also developed three ontologies: section header ontology from extracted section headers using RDF standard,

concept ontology comprising entities representing five quality metrics from SNOMED (Diabetes, Hypertension, Cardiac Surgery, Transient Ischemic Attack, CNS tumor), and a clinical note ontology that represented clinical note elements and their relationships. All ontologies (Web Ontology Language format) and patient notes (RDFs) were imported into a triple store (AllegroGraph®) as classes and instances respectively. SPARQL information retrieval protocol was used for reporting extracted concepts under four settings: base Natural Language Processing (NLP) output, inclusion of concept ontology, exclusion of negated concepts, and inclusion of section header ontology. Existing manual abstraction data from surgical clinical reviewers, on the same set of patients and documents, was considered as the gold standard.

Micro-average results of statistical agreement tests on the base NLP output showed an increase from 59%, 81%, and 68% to 74%, 91%, and 82% (Precision, Recall, F-Measure) respectively after incremental addition of ontology layers.

Our study introduced a framework that may contribute to advances in “complementary” components for the existing information extraction systems. The application of an ontology-based approach for natural language processing in our study has provided mechanisms for increasing the performance of such tools. The pivot point for extracting more meaningful quality metrics from clinical narratives is the abstraction of contextual semantics hidden in the notes. We have defined some of these semantics and quantified them in multiple complementary layers in order to demonstrate the importance and

applicability of an ontology-based approach in quality metric extraction. The application of such ontology layers introduces powerful new ways of querying context dependent entities from clinical texts.

Rigorous evaluation is still necessary to ensure the quality of these “complementary” NLP systems. Moreover, research is needed for creating and updating evaluation guidelines and criteria for assessment of performance and efficiency of ontology-based information extraction in healthcare and to provide a consistent baseline for the purpose of comparing alternative approaches.

Vita

1998.....MD, Medicine, Shahid Beheshti University
of Medical Sciences, Tehran Iran

1998.....BA, Foreign Language and Translation
(English), Azad University, Tehran, Iran

2004-2007MS, Health Informatics, University of
Missouri-Columbia, MO

2005 to 2008Structured Clinical Documentation &
Interface Terminology, Vanderbilt University Medical Center, Nashville, TN

2008 to present.....Clinical Data Modeling, Natural Language
Processing & Ontology, MD Anderson Cancer Center, Houston, TX

2009 to present.....PhD Candidate, Biomedical Informatics,
University of Texas Health Science Center, Houston, TX

2011.....Fellowship Award, National Human Genome Institute

2012..... Fellowship Award, Agency for Healthcare Research & Quality (AHRQ)

2013..... Fellowship Award, Agency for Healthcare Research & Quality (AHRQ)

Field of Study

Biomedical Informatics

Table of Contents

Dedication	ii
Acknowledgements	iii
Abstract	iv
Vita.....	vii
Table of Contents	viii
List of Tables	x
List of Figures	xii
Chapter 1: Introduction	1
Chapter 2: Literature Review	4
I. Quality improvement	4
Why measure quality?.....	5
Sources and types of quality metrics	8
Problems in modeling and extraction of quality metrics	10
II. Information extraction.....	12
Natural language processing (NLP) theories	15
Theories of meaning: Semantic theories.....	19
Clinical Narratives	22
Terminologies and NLP Knowledgebase	24
Evaluation of NLP in Healthcare Informatics	27
Ontologies in healthcare	32
Existing work in clinical information extraction using ontologies	34
Chapter 3: Methodology	38
Building Ontologies	40
Ontology Language & Editing Environment	40
Patient Selection.....	41
Gold Standard	42
Metric selection.....	44

Patient note extraction & pre-processing.....	44
Natural Language Processing (NLP) Engine.....	46
Data format and repository type	49
Evaluation of Ontologies	50
I - Formal Methods	51
II – Statistical Agreement Tests.....	52
Chapter 4: Findings.....	58
Transcribed documents	58
Ontologies	60
<i>Section Header Ontology</i>	60
<i>Quality Metric Ontology</i>	62
<i>Clinical Note Ontology</i>	64
Evaluation of quality metric extraction.....	66
Error Analysis	72
Limitations	74
MetaMap	75
Chapter 5: Conclusions, Discussions, and Future Directions	77
Conclusions.....	79
Future Directions	80
References	83
Appendix A: Diabetic Medication Nursing Reference	93

List of Tables

Table 1 <i>Examples of organizations and measurement collection programs</i>	6
Table 2 – <i>Selected NSQIP quality metrics reported values</i>	44
Table 3 – <i>Contingency table</i>	54
Table 4 - <i>Micro-averaging multiple contingency tables</i>	57
Table 5 - <i>Macro-averaging multiple contingency tables</i>	57
Table 6 - <i>Top 20 Transcribed Patient Notes types and their frequencies</i>	58
Table 7 - <i>Selected 8 note type frequencies and section header counts</i>	59
Table 8 - <i>Automatic section extraction performance compared to a gold standard</i>	61
Table 9 - <i>Section header distribution within 5 selected quality metrics</i>	61
Table 10 – <i>Number of concepts included in quality metric ontology</i>	63
Table 11 - <i>Instance count of the main patient note ontology objects</i>	65
Table 12 - <i>Agreements statistics results after addition of each layer (cumulative) for the quality metrics extracted from narrative texts</i>	67
Table 13 - <i>Micro-averaging the results of all 5 quality metrics combined</i>	67
Table 14 - <i>Macro-averaging combined result of agreement tests for 5 quality metrics...</i>	68
Table 15 - <i>Agreements statistics for each quality metric extracted from narrative texts. The difference is calculated for each layer in isolation and relative to the base NLP output</i>	70

Table 16 - <i>Micro-average result of non-cumulative agreement tests for the five quality metrics under study</i>	71
Table 17 - <i>Macro-average result of non-cumulative agreement tests for the five quality metrics under study</i>	71
Table 18 – <i>Source of discrepancies in false positive and false negative cases</i>	72

List of Figures

Figure 1- <i>The relationship between content and context</i>	20
Figure 2 - <i>The relationship between character, context, content, and circumstance</i>	21
Figure 3- <i>MetaMap architecture</i>	27
Figure 4- <i>Schematic view of the proposed ontology-based quality metric extraction framework</i>	39
Figure 5 – <i>Patient notes processing pipeline: Extraction and pre-processing</i>	45
Figure 6 - <i>Patient notes processing pipeline: Conversion of processed notes to XML</i>	45
Figure 7 - <i>Parsed and regionized patient note are converted into XML format</i>	47
Figure 8 - <i>Expanded view of MetaMap node (MMO) in the XML output</i>	48
Figure 9 - <i>Conversion of a processed note into RDF and extraction of a quality metric</i> .	49
Figure 10 - <i>Use of subject matter expert for comparison of the data generated by a system</i>	53
Figure 11 - <i>Distribution of note counts per patient</i>	60
Figure 12 - <i>Section header classification and synonym assignment using SKOS</i>	62
Figure 13 - <i>Diabetes Mellitus ontology hierarchy</i>	63
Figure 14 - <i>Patient note ontology: Objects are shown in gold, objects properties in blue, and data type properties in green</i>	64
Figure 15 - <i>Sample query from instances populated in the patient note ontology</i>	66

Figure 16 - <i>Micro-average combined result of agreement tests for the five quality metrics</i>	
.....	68
Figure 17 - <i>Macro-averaging combined agreement tests for 5 quality metrics</i>	69
Figure 18 – <i>The difference in F measure for each layer relative to the base NLP output</i>	71

Chapter 1: Introduction

The Institute of Medicine reports a growing demand in recent years for quality improvement within the healthcare industry (Committee on Identifying and Preventing Medication Errors. Institute of Medicine., 2006; Committee on Quality of Health Care in America. Institute of Medicine., 2000, 2001; Committee on Redesigning Health Insurance Performance Measures Payment and Performance Improvement Programs. Institute of Medicine., 2007). In response, numerous organizations have been involved in the development and reporting of quality measurement metrics. However, the quality metrics development process is subjective in nature (Miller, 2010) and competing interests exist among stakeholders. As a result, conflicting data definitions from different sources shift the burden of accurate and reliable metrics extraction and reporting to the healthcare providers (Committee on Redesigning Health Insurance Performance Measures Payment and Performance Improvement Programs. Institute of Medicine., 2006; Tang, Ralston, Arrigotti, Qureshi, & Graham, 2007; Velamuri, 2010). Furthermore, manual abstraction of quality metrics (Leavitt, 2008; Velamuri, 2010), diverse implementation of Electronic Health Record (EHR) Systems (McDonald, 1997; Velamuri, 2010), and the lack of standards for integration across disparate clinical and research data sources (Chong, Marwadi, Supekar, & Lee, 2003) deepens the complexity of consistent, valid, explicit, and comparable quality measurement extraction and reporting tasks within healthcare provider organizations.

In order to construct a quality metric extraction framework, based on standards, concepts should be defined explicitly, such that heterogeneous information from different sources can be reliably mapped and compared based on those concepts. According to the theories of meaning, semantics can define the explicit meaning of an entity relative to the content, context, and state in which the entity is expressed. The real meaning of entities can then be used for formal definition, disambiguation, and conceptual modeling in a given domain of discourse. While a reference information model, like the proposed National Quality Forum Data Model ("National Quality Forum Quality Data Model,") or eMeasures (Velamuri, 2010), can be used for deriving a syntactic data model (Carlson, Farkash, & Timm, 2010), it does not represent such a shared and comparable data semantics (Smith & Ceusters, 2006) for harmonized representation of heterogeneous schemas (Bianchi et al., 2009; Carlson et al., 2010). In addition, neither is there a well-defined interface between such information models and the EHR systems (Ferranti, Musser, Kawamoto, & Hammond, 2006) nor can quality metrics be represented solely by such complex standards (E. Muir; Eliot Muir, 2013; Pisanelli & Gangemi, 2004). Hence, quality metrics developed by diverse organizations, as well as provider's internal metrics, cannot be modeled exclusively, compared explicitly, and *extracted* unambiguously by reference standard information models alone (Eliot Muir, 2013). Therefore, we propose an ontological extraction framework with clear semantics to overcome such shortcomings.

In the first phase of this study we will explore existing quality measurement metrics and their components and derive a comprehensive conceptual model using a standard

terminology and semantic specification (McGuinness & Van Harmelen, 2004b). We intend to use formal and concept extraction methods to construct and extend an unambiguous semantic nomenclature from the explored components. The methods will explicitly define all concepts, show relationships among concepts and their contexts, normalize attributes, binds concepts into standard terminologies (Bianchi et al., 2009), and facilitate query functionalities (Kamal, Borlawsky, & Payne, 2007).

In the second phase we will perform a series of federated queries, using multiple ontological layers, on a target group of patient notes and compare the results against the current manual abstraction data for the purpose of functional validation of the model. Domain experts will validate completeness, domain coverage, and accuracy of the model. Existing conventional semantic rule engines will be used for structural validation of the model.

The host institution for this study, MD Anderson Cancer Center, is the largest freestanding cancer center in the world. There were 115,000 patients who visited MD Anderson in 2012 ("Facts and History - Quick Facts 2013 | MD Anderson Cancer Center," 2013) , thus providing the primary investigator with a large amount of patient data for validation and applicability of the proposed framework.

Chapter 2: Literature Review

In the following sections we will briefly review the quality improvement process and issues related to quality measurement in healthcare followed by an overview of Natural Language Processing (NLP) and semantic theories. The first section outlines a brief history of quality measurement and its importance in healthcare. In addition, types and sources of quality metrics are explored and current challenges in extraction and reporting of metrics are discussed. In the second section, we briefly review the theoretical background of NLP and theories of semantics and meaning, specifically the propositional semantic theory, and its application in conceptual modeling and extraction of quality metrics.

I. Quality improvement

In 1920s and 1930s, Shewhartf (Shewhart, 1931), Deming (Deming, 2000), and Juran (Juran, 2004) introduced the initial concept of Quality Assurance. Quality assurance itself consists of core activities such as quality definition, quality measurement, and quality improvement (Quality Assurance Project., 2001). In 1966, Donabedian (A. Donabedian, 1966) defined a framework for quality measurement in the healthcare industry and described three major components in his framework: structure, process, and outcome. Structure measures refer to all resources, including infrastructures, technologies, and systems that are required for a given process of care. All procedures performed

on patients, including but not limited to diagnostic and therapeutic procedures, are measured by Process outcomes. Procedure outcomes during patients' care processes are captured and represented by Outcome measures (Avedis Donabedian, 1980).

The Institute of Medicine (IOM) included several key concepts in its definition of Quality of Care in 1990: *"Quality of care is the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge"* (Committee to Design a Strategy for Quality Review and Assurance in Medicare. Institute of Medicine., 1990). According to this definition, improving the quality of care applies to all domains of healthcare services, from preventive to palliative and from acute to chronic, and at both individual and population levels. The definition also emphasizes the Donabedian framework and the fact that providing optimal processes of care alone may not necessarily result in excellent patient outcomes and vice versa. Finally, knowledge management in the form of constant knowledge acquisition, revision, and sharing plays an important role in achieving the state of high quality of care (Chassin & Galvin, 1998).

The Institute of Medicine also acknowledges that effective use of information technology in clinical information systems, for automating the quality measurement collection process, is among healthcare organizations' top challenges for improving quality of care (Committee on Quality of Health Care in America. Institute of Medicine., 2001).

Why measure quality?

The Institute of Medicine reports a growing demand in recent years for quality improvement within the healthcare industry (Committee on Identifying and Preventing

Medication Errors. Institute of Medicine., 2006; Committee on Quality of Health Care in America. Institute of Medicine., 2000, 2001; Committee on Redesigning Health Insurance Performance Measures Payment and Performance Improvement Programs. Institute of Medicine., 2007). In response, numerous governmental agencies, consortiums, hospital accreditation groups, and private organizations are involved in the development and collection process of metrics (Kavanagh, Adams, & Wang, 2009; National Committee for Quality Assurance.) (Table 1).

Table 1 *Examples of organizations and measurement collection programs*

Quality Development Organizations	Metric Collection Programs
Agency for Health Research and Quality (AHRQ)	Agency for Health Research and Quality, Quality Indicator (QI)
Centers for Disease Control and Prevention (CDC)	Consumer Assessment of Healthcare Providers and Systems (CAHPS)
Centers for Medicare & Medicaid Services (CMS)	CMS Core
National Cancer Institute (NCI)	Healthcare Effectiveness Data and Information Set (HEDIS)
Comprehensive Cancer Care Consortium Quality Improvement (C4QI)	Hospital Outpatient Quality Data Reporting Program Support Contractor (HOP QDRP)
National Comprehensive Cancer Network (NCCN)	National Comprehensive Cancer Network Guidelines
University Health System Consortium (UHC)	National Surgical Quality Improvement Program (NSQIP)
Institute of Medicine (IOM)	Physician Quality Reporting Initiative (PQRI)
National Quality Forum (NQF)	National Quality Forum Data Model
Institute for Clinical System Improvement (ICSI)	American College of Surgeons Quality Collaboration (ACS QC)
Physician Consortium for Performance Improvement (PCPI)	Physician Consortium for Performance Improvement Measures
American College of Surgeons (ACoS)	Quality Oncology Practice Initiative (QOPI)
American Nurses Association (ANA)	National Database of Nursing Quality Indicators (NDNQI)
American Society of Clinical Oncology (ASCO)	National Quality Measures Clearinghouse (NQMC)
Institute of Healthcare Improvement (IHI)	Outcome-based Quality Improvement (OBQI)
Leapfrog	Press Ganey

Quality measurement development is an iterative and often lengthy process (National Committee for Quality Assurance.). It starts with the acquisition of evidence-based and subject matter expert knowledge in a selected domain of care. Several techniques such as consensus rating, Delphi technique (Hasson, Keeney, & McKenna, 2000), and the RAND appropriateness method (Brook et al., 1991) are being used by both private and public organizations for defining and building quality measurements. The development process continues with the selection of panel members, scientific literature review, metric candidacy, and evaluation sessions by representatives from all involved stakeholders. The process ends with validating applicability of the candidate metric by a field test implementation (Kavanagh et al., 2009).

In order to improve quality metrics they should be measured first. Such measurement facilitates defining best practices in a given domain of care, identifying and comparing variation of care, creating a foundation for structural definition of quality improvement, and evaluating treatment and procedure effects (Kavanagh et al., 2009). Other important reasons for measuring quality in healthcare include making knowledgeable decisions from existing choices by healthcare consumers and purchasers, selecting appropriate treatment for patients, and doing well-informed referrals for providers (Hewitt & Simone, 1999). In addition, many studies and reports have shown the benefits of quality of care improvement in terms of saved costs and lives (Chassin et al., 1987; Committee on Quality of Health Care in America. Institute of Medicine., 2000; National Committee for Quality Assurance.; Thomas et al., 1999).

The well-known report from Institute of Medicine, claiming 44,000 to 98,000 deaths due to medical errors (Committee on Quality of Health Care in America. Institute of Medicine., 2000), shook the healthcare community in 1999. In another study by Thomas (Thomas et al., 1999), the total national cost of preventable adverse events was estimated between \$17 and \$29 billion of which healthcare related costs constitute \$18 billion of the total. Recent data from the Healthcare Effectiveness Data and Information Set (HEDIS) also showed between \$4.5 to \$7.4 billion dollars and 50,000 to 186,000 lives per year can be saved in the United States by improving only 75 quality measures across 8 domains of care (National Committee for Quality Assurance.).

Continuous measurement of quality metrics and learning from medical mishaps can also help reduce preventable harms due to medical errors. Such harmful events are generally categorized as underuse, overuse and misuse of medical treatments. In two separate studies, Chassin et al. showed under-usage of beta blockers were accountable for 18,000 loss of lives (Chassin, 1997) and over-usage of endoscopic and angiographic procedures responsible for 17% of excessive usage of therapeutic procedures each year in the United States (Chassin et al., 1987)

Sources and types of quality metrics

Upon selection of a list of quality metrics for reporting, the extraction phase begins by identifying the sources of the metrics. The most prevalent sources are dictated notes and claims data. However, manual extraction of the data from transcribed notes (abstraction) is both time consuming and costly. Furthermore, claims data is not considered as a reliable source of information. Billing and administrative (or claims) data only shows the

pattern of healthcare resource utilization, if coded appropriately; hence, it does not accurately reflect the care provided to the patients (Kavanagh et al., 2009; McGlynn et al., 2003).

Satisfaction surveys handed over to the patients during their visit is another source for quality metrics data. Although these surveys reasonably echo patient's perception of quality of care; capturing and processing such information is a time consuming and costly task for provider organizations. National registries at the local, state, national, and international levels, like Surveillance Epidemiology and End Results (SEER) (Ries, 1999) and National Cancer Data Base (NCDB) (Raval, Bilimoria, Stewart, Bentrem, & Ko, 2009) are examples of other sources of information for quality metrics that are mostly used for epidemiological studies.

EHR systems are considered the best source for extracting quality metrics because they contain longitudinal information of patients that accurately reflects the actual process of care. However, a gap between health information technology and extraction of structured and unstructured information from EHR systems still remains largely open within healthcare organizations (Hewitt & Simone, 1999; Velamuri, 2010).

While acquiring outcome measurement from administrative and billing data from EHR systems seems to be much easier than process measurements, the value of analyzing process metrics is much higher than outcome measurements collection programs (Kavanagh et al., 2009). In addition, process measures are more sensitive to manipulation and can be easily controlled during a given healthcare process. On the contrary, outcome measures can be affected by contributing factors such as patient pre-conditions, severity

of the disease, and environmental factors, therefore, undesirable outcomes do not necessarily correlate to poor quality of care or vice versa (Kavanagh et al., 2009). In order to achieve a high quality level of the patient care process, ideally, quality measurement should be done on all the three components of Donabedian framework. However, due to the complex nature of relationship among these components and the difficulties in extracting required data, a comprehensive information management solution that can capture information from all three components is not available today. Defining such a relationship between process and outcome measures as well as providing unambiguous and clear definition of target population, setting, time frame, and metric components remain among the desiderata for measurement indicators (Kavanagh et al., 2009).

Problems in modeling and extraction of quality metrics

The meaning of Quality in healthcare is vague and a standard definition of the term "*Quality*" health care is still lacking. The confusion and multiple languages around quality measurement comes from the fact that involved entities in healthcare systems translate their interests into their own terms for defining quality measurements, hence, making a standard definition for a given quality measurement quite difficult. Any standard definition, therefore, should disambiguate models and the terminologies used in those models in the domain of quality (Saturno, 1999).

A model is an abstract representation of a real world entity. Many models can be drafted from the same object of which none could be labeled as the most complete. Also, models have a tendency to degrade over time (Coiera, 2003). In the health care industry, there are

many models for quality improvement and the better we can fit model attributes into the real world representation the better we can explain and expect the outcomes. As it is true for every model, existing quality models in health care should not only be updated constantly, to better represent the reality of healthcare system, but also be detailed enough about every aspect of the quality of health care. Therefore, an inclusive approach that captures multiple views from all existing quality models is more desirable than adopting an exclusive one (Coiera, 2003; Quality Assurance Project., 2001).

An ideal model of quality of care should provide a comprehensive “360 degree” view for all stakeholders including patients, physicians, health plans, public health officials, and policy makers (Spinks et al., 2011; Wimmer, Scholl, & Grönlund, 2007). Each stakeholder has its own interpretation of quality and views the model from a different angle (Weng, Gennari, & Fridsma, 2007). For example, a patient view of quality is usually interpreted as responsiveness of the provider (in terms of speed and timeliness), expected mortality, and available alternative choices of care. On the other hand, providers look at the quality of care from the perspective of most excellent outcome based on their clinical judgments. Healthcare regulators consider appropriateness of clinical interventions and outcomes whereas public health officials usually look for epidemiological data such as mortality and morbidity. Therefore, in order to construct such an overarching model, quality metric concepts should be defined explicitly, such that heterogeneous information from different sources can be reliably mapped and compared based on those concepts.

Standard models for quality measurements, like the National Quality Forum (NQF) Data Model ("National Quality Forum Quality Data Model,"), have been proposed recently from measurement development and endorsement organizations to provide a standard conceptualization of quality metrics to be consumed by all stakeholders. While a reference information model, like the proposed NQF Data Model ("National Quality Forum Quality Data Model,"), or eMeasures (Velamuri, 2010), can be used for deriving a syntactic data model (Carlson et al., 2010), it does not represent such a shared and comparable data semantics (Smith & Ceusters, 2006) for harmonized representation of heterogeneous schemas (Bianchi et al., 2009; Carlson et al., 2010). In addition, neither is there a well-defined interface between such information models and EHR systems (Ferranti et al., 2006) nor can cancer quality metrics be represented solely by such a complex syntactical standard (E. Muir; Eliot Muir, 2013; Pisanelli & Gangemi, 2004). Hence, quality metrics developed by diverse organizations, as well as provider's internal metrics, cannot be modeled exclusively, compared explicitly, and extracted unambiguously by reference standards such as the proposed reference information models (Eliot Muir, 2013).

II. Information extraction

Information extraction systems have been developed and in use for the past half a century. The main driver for development of such systems was laid out during Message Understanding Conferences (MUC) hosted by Defense Advanced Research Project Agency (DARPA) between 1987 and 1998. Entity recognition and relation extraction were the focus of those conferences, which later led to the development of the first

information extraction systems in late 1970s by DeJong for Reuter Company. Information extraction systems typically perform one or a combination of these tasks: Natural Language Processing (NLP), Named Entity Recognition, Text Mining, and Information Retrieval. All of these tasks have several applications in health care domain. For example, NLP is typically used for concept recognition from narrative texts (like signs & symptoms, disorders, medications), detection of relevant documents, summarizing patient information, acquisition of new knowledge, validation of existing knowledge, and data integration among disparate sources of data.

A number of NLP systems have been developed and utilized in the medical domain. These systems have focused on areas such as clinical decision support, quality metrics reporting, and patient data management. Other types of NLP systems, originally developed outside the medical field, have also been employed for concepts such as automated encoding, literature indexing and vocabulary development (C. Friedman & Hripcsak, 1998). The first clinical NLP systems were developed around 1986 by researchers at the New York University and were referred to as the Linguistic String Project – Medical Language Processor (LSP-MLP) (Sager, Lyman, Nhan, & Tick, 1995). It is considered the founding father of subsequent clinical NLP systems and aimed at extraction of patient signs & symptoms and medication related information. From information retrieval's perspective LSP-MLP reached a precision & recall of 98.6% and 92.5% respectively.

The Specialist NLP tools have been developed by The Lexical System Group of the Lister Hill National Center for Biomedical Communication (McCray & Nelson, 1995) to

facilitate interactions between user's and biomedical information languages. For every given entity in the dictionary, a semantic type has been assigned and all semantic types are also connected to each other through the Semantic Network (Humphreys & Lindberg, 1992). Currently, 133 semantic types exist in the Semantic Network and 54 semantic relations connect all of them through a predefined hierarchy ("Semantic Network," 2009). Due to the lack of standard written grammar in clinical narratives, Center d' Informatique Hospitaliere of the Hopital Cantonal de Geneve adopted a different approach (Proximity Processing) for processing of clinical texts (R. H. Baud, Rassinoux, & Scherrer, 1992; Scherrer, Revillard, Borst, Berthoud, & Lovis, 1994). The system was called Representation du Contenu Informationnel des Textes medicaux (RECIT) and its logic was based on the fact that it is highly probable that one word becomes the modifier of another word when those two words occur together. This approach was less language dependent and emphasized more on the semantics of the narrative text than syntax (A. Rassinoux, Baud, & Scherrer, 1990; A. M. Rassinoux, Michel, Juge, Baud, & Scherrer, 1994).

Carol Friedman in the Columbia University of New York developed one of the popular NLP systems in medicine in 1993. This NLP engine is called MEDical Language Extraction and Encoding System (MEDLEE) and has been widely used by the academic community (C. Friedman, Cimino, & Johnson, 1993). MEDLEE was originally tested on only radiology reports and discharge summaries but later on was extended to other clinical note types. The primary driver for MELEE is semantic rules, however, syntactic grammar has also been incorporated in order to increase efficacy of the system.

MEDLEE reached a sensitivity and specificity level of 81% and 98% respectively (G. Hripcsak et al., 1995). There are also many other non-English NLP systems developed for German, French, and Japanese speaking users; Aristoe, Rime, Meditas, Metexas and Medi-cat (Peter Spyns, 2000), to name a few. In all NLP systems, the developers were targeting two fundamental tasks: language analysis and knowledge representation. Based on the amount of focus on either of these tasks in each approach, the level of language and domain dependencies varies during text processing. In a language independent/domain dependent approach, knowledge engineering and domain modeling are essential for information extraction. This approach requires more human interaction in order to build and create domain knowledge models in order to “guide” or “compliment” the NLP system and “infer” the meaning and extract information from text. On the other hand, in the language dependent/domain independent approach, typical sentence parsers are employed and the output of syntactical full parsers is fed into a semantic processor for further analysis (P. Spyns, 1996).

Natural language processing (NLP) theories

NLP is “an automated technique that converts narrative documents into a coded form that is appropriate for computer-based analysis” (Melton & Hripcsak, 2005). Carbonell and Hayes (Shapiro, 1992), in the Encyclopedia of Artificial Intelligence in 1992, defined NLP as “the formulation and investigation of computationally effective mechanism for communication through natural language”. The objective of designing such processing systems is for computers to understand the “language” of humans. In other words, the basis of NLP lies in modeling language as a form of communication, in which one human

(sender) emits a message represented by a set of specific acoustic or graphic signs to another person (receiver). Obviously, in order for the receiver to understand the message, the sender and receiver need to share some common sense knowledge (P. Spyns, 1996). Charles Morris explains the concept using 'syntactics- semantics- pragmatics' triplet which has become the cornerstone of NLP. Pragmatics represents the complete environment of the sender or receiver, semantics is the relationship of expressions to their meaning, and syntactics is the study of approaches that construct compound signs from smaller parts (Morris, 1971).

While earlier work can be found, the history of NLP is said to start with Alan Turing and his paper proposing what is now called the Turing test as a criterion of intelligence. The criterion calls for a computer program impersonating a human sender in real-time, such that a human receiver cannot tell the difference with a real human sender based on the conversation alone (Turing, 1950). There are multiple approaches to NLP, some examples are the symbolic approaches – where knowledge about language is encoded in various representational formats; NL Analysis – which runs through lexical analysis, syntactic, semantic, and pragmatic analysis; NL Generation – which is employed to generate fluent text from underlying information; and finally, empirical approaches, which include statistical analysis on large amount of data. Part of Speech (POS) tagging, alignment, collocations, and word-sense disambiguation are some example tasks from the empirical approach to NLP (Dale, Moisl, & Somers, 2000).

Current NLP applications in healthcare can be categorized into two groups; statistical and linguistical. All classical machine learning and statistical tools can be used in the

statistical approach, hence, make it a good method for fast text classification purposes (like Google). On the other hand, a linguistic or symbolic NLP approach is usually used for meaning (knowledge) extraction by incorporating shallow or chunking parsers (tokenization). Most of the clinical information systems, however, use a combination approach of both statistical and linguistic methods. Nevertheless, NLP systems can also be categorized in other ways; some experts classify them into partial and complete systems according to the level of morphologic (word), conceptual (semantic), and sentence level (knowledge representation) processing (P. Spyns, 1996).

One strong consensus in the field of linguistics at this time is to restrict research to well-defined sub-languages – i.e. a technical language that is used by the various actors in the technical field to pass specific messages. This technical language has some main differences from the general language. First, in a technical language a considerable amount of general language words can take a more restricted and specific meaning; and there also exists a very specific vocabulary that is almost exclusively used in that domain. Second, the sentence construction rules for the technical vocabulary are also different; omission of words, that are not strictly necessary, creates a telegraphic style seen commonly in clinical notes. Finally, every sub-language has its own idiosyncratic expressions, which are very difficult to understand when used outside of the medical domain, since they are created for a concise description of patterns in their respective territory. On the plus side, technical languages are not as flexible as the general language; for example a patient discharge summary does not contain verbs in second person or questions. Names are not mentioned as often as other types of text, and the entities that

construct the context are usually known (patient, provider, facility, etc.). Formulation and evaluation are also facilitated by the fact that NLP systems for healthcare share a common set of objectives, such as improving patient care or facilitating the work of the provider (C. Friedman & Hripcsak, 1998). This makes it easier to write tools for these subsets (P. Spyns, 1996)

For the field of medicine, the 'medical jargon' which is composed of many terms with Latin or Greek parts gives some universality to the medical notations. Moreover, the practice of medicine is very similar in different parts of the world, and for a medical message, since the sender and receiver are both doctors, they share an amount of medical knowledge that need not be addressed explicitly in the message. These characteristics favor the application of artificial intelligence (AI) for NLP in healthcare. AI constructs (frames, scripts, and domain modeling) allow for deductive and temporal reasoning, inference, coreferentiality, and reference resolution; and specific theories, such as Discourse Representation Theory, show promise in coping with such problems in the processing of medical text because of such properties (P. Spyns, 1996).

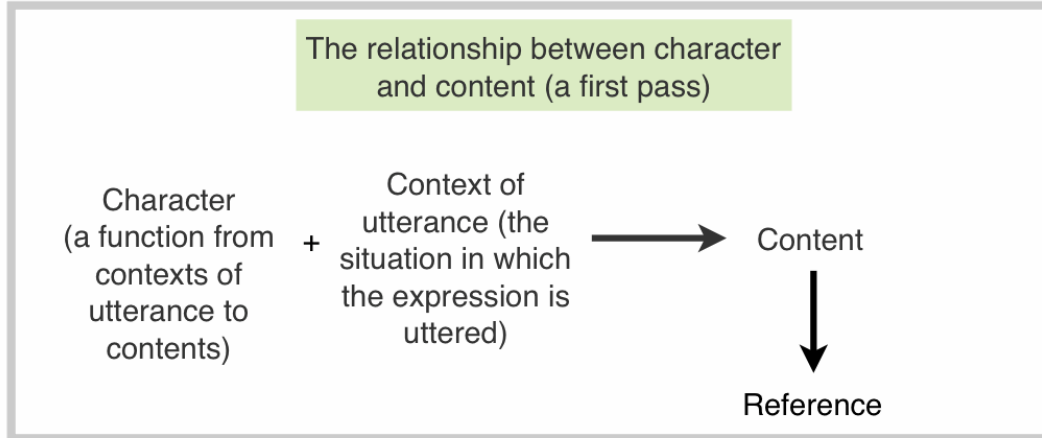
On the other hand, in clinical medicine each domain sub-language is a technical language used by domain expert in a given domain of care and it could be semantic or syntactic specific or a combination of both. Some of the normal words that are used in these sub-languages have the same meaning in other non-related domains. However, many other normal words in clinical sub-language become context dependent and have a different meaning (A. M. Rassinoux et al., 1995). For example “history” is usually interpreted as “patient medical or surgical history” or “Previous Myocardial Infarction history” can be

applied to the patient or a family of a patient based on the location (section header) where it is found in a typical clinical narrative (Denny, Miller, Johnson, & Spickard III, 2008). The grammar seen in clinical transcribed documents is often poor due to non-standard ambiguous abbreviation usage, inferred concepts, and poorly segmented sentences. Meta data about notes are usually missing or incomplete. These properties result in a high degree of dynamic semantics in clinical context (A. M. Rassinoux et al., 1995) which we will review it in the next section.

Theories of meaning: Semantic theories

The theories of meaning have been the center of many philosophical debates for the past fifty years. There are two main categories of theories around meaning; semantic and foundational theories. Semantics theories focus on the meaning of expressions (entities), types of expression (classification), and assignment of semantic symbols to the expressions (specification of the meaning) whereas, in the foundational theories explanations and descriptions are conveyed for the sociological and psychological facts about expressions (Lewis, 1972). Semantic theories can further be classified into propositional and non-propositional types but for the sake of our discussion we will only focus on propositional semantic theories and describe their major elements (reference, content, context, circumstance) and the relationships that exists among them (Figure 1).

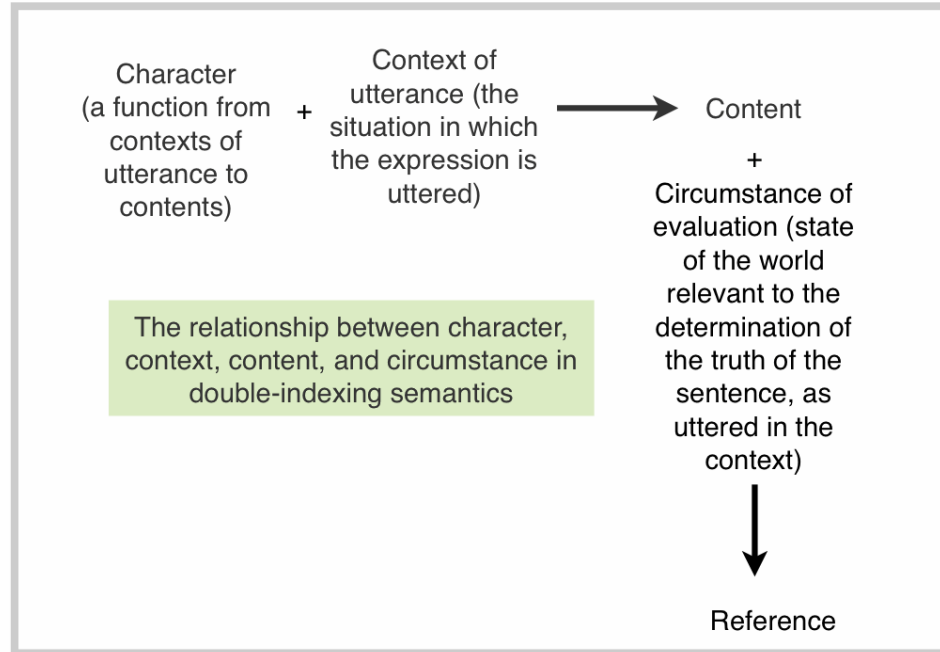
Figure 1- *The relationship between content and context*



Speaks, Jeff, "Theories of Meaning", The Stanford Encyclopedia of Philosophy. Summer 2011 Edition

According to semantic theories, expressions are paired with values. Such values are sometimes called entity or reference. However, the value of an expression is subject to change based on the situation (context) in which the expression occurs. Such conditional expressions are often called indexical or context dependent expressions (Kaplan, 1979). We should not confuse the meaning of an expression as whether attributable to its content or character but rather to think that both elements participate in the meaning of an expression with a known context. Nevertheless, context is not the only determining factor in discovering the real (or explicit) meaning of an expression. Depending on the state or circumstance in which the expression occurs or is being evaluated, the meaning of an expression could be subject to a second indexical change. Such expressions require double indexing and become semantically both context and circumstance dependent (Speaks, 2006) (Figure 2).

Figure 2 - *The relationship between character, context, content, and circumstance*



Speaks, Jeff, "Theories of Meaning", The Stanford Encyclopedia of Philosophy. Summer 2011 Edition

In many instances, the context and circumstance of an expression are the same and additional indexing, for the purpose of clarity, is not required. In some other occasions, the context of an expression is included in the content (pre-coordination) and indexical approaches become redundant. So, the *real meaning* of an entity becomes *relative* to the content, context, and state in which the entity is expressed and, therefore, can be represented (or modeled) in different ways. Identification of such representational variations in expressions (especially in clinical expressions) and providing equivalencies among such representation is a crucial task in any knowledge modeling and information management activities.

Discourse Representation Theory (DRT), introduced by Hans Kamp in 1981, is one of the theories of 'dynamic semantics'. These theories focus on context dependence of meaning; that is, by accepting that utterances in natural language are only meaningful if their context is taken into account and observing that each utterance in turn contributes to the context in which it is made. Dynamic semantics asserts that each utterance will change the context into a new context that in turn influences the interpretation of whatever utterance comes next. In this perspective, contrary to the other classical conception of formal semantics, the meaning of a sentence would be its capacity to change the underlying context, not its truth conditions as is the case with static semantics. DRT differs from other theories of dynamic semantics in that it still attributes a prominent role to truth conditions, so much so that some authors have classified it as static (Kamp, Van Genabith, & Reyle, 2011).

Nevertheless, DRT still meets all the criteria that define the basis of a dynamic semantics theory. DRT enhances the machinery of formal semantics to provide the capability of capturing the cohesion between sentences in a given text. Much of this cohesion comes from the anaphoric properties of natural language (i.e. the ability of each expression to refer to other expressions in text). As example, Pronominal forms such as she, he, him, her, and it as well as tense are anaphoric devices because they enable a sentence to refer to specific concepts in the other parts of the text (Geurts & Beaver, 2011).

Clinical Narratives

During the past decade, the healthcare service providers have shown a substantial interest in Electronic Medical Record (EMR) systems. As new systems are developed and

deployed, the use of such system increases in the healthcare industry (Wager, Lee, & Glaser, 2009). These systems are supposed to be superior to paper based records in many respects, such as accessibility, readability, accuracy, and more importantly availability of data. Traditionally, physicians used patient records as a memory support for future encounters with the patient. With the introduction of Healthcare Information Technology (HIT) systems such as EMR systems, the informatics use cases for patient records has shifted to areas such as quality measurement, decision support, and data integration. This paradigm shift has brought new requirements with respect to the content, structure, and accuracy of information contained in a patient record (W Ceusters, Lovis, Rector, & Baud, 1996).

As such, natural language does not meet the criteria to explicitly and unambiguously extract the important information that is entered in patient records in a manner that is fit for computer analysis. Nevertheless, natural language is still easier, more expressive, and more frequently used to transmit complex information about patients (Scherrer et al., 1994) which accounts for the popularity and market for solutions that capture it for medical records such as handwriting or speech recognition systems. Physicians typically interact with such systems on a daily basis in order to record and retrieve patient information. As a general rule, in a typical patient care environment, patient information is “dictated” by physician and then “transcribed” and stored into the EMR system in free text (or narrative) format ("Medical Records, Coding & Health Information Management: AHIMA Facts," 2013; Milewski, Govindaraju, & Bhardwaj, 2009). Some research interest has therefore focused on extraction of information from narrative unstructured

texts by Natural Language Processing (NLP) engines, though it still remains as one of most challenging tasks in biomedical informatics domain (A. M. Rassinoux et al., 1994).

Terminologies and NLP Knowledgebase

Since the number of clinical concepts, including their synonyms, is rather large (over 100,000), it takes a huge amount of effort to create a comprehensible terminology. The National Library of Medicine (NLM) released Unified Medical Language System (UMLS) in 1986 order to provide a collection of terminologies in the biomedical domain (Carol Friedman & Hripcsak, 1999). In UMLS, each concept is given a unique identifier (Concept Unique Identifier or CUI) and all synonymous terms are associated with that CUI. Such coding can be used in NLP systems to map synonymous phrases in text to standard terminologies, as UMLS contains concepts from a variety of sources. The semantic network within UMLS system assigns semantic types to the concepts. For example, Neoplastic Process is a type of Disease or Syndrome class ("UMLS® Reference Manual," 2009). Assignment of semantic types to the concepts empowers NLP systems to link concepts with appropriate relationships. The SPECIALIST lexicon, can be used in NLP extraction, indexing, and terminology development activities (Carol Friedman & Hripcsak, 1999). Other Nomenclatures are also important knowledge sources. SNOMED CT (Zweigenbaum & Courtois, 1998) and ICD10 (R. Baud, Lovis, Rassinoux, Michel, & Scherrer, 1997) have been used as knowledge sources, since both are particularly useful in settings where multilingual text needs to be processed. These terminologies are included in the current release of UMLS ("UMLS® Reference Manual," 2009).

With the increasing number of terminologies becoming available in medicine, their

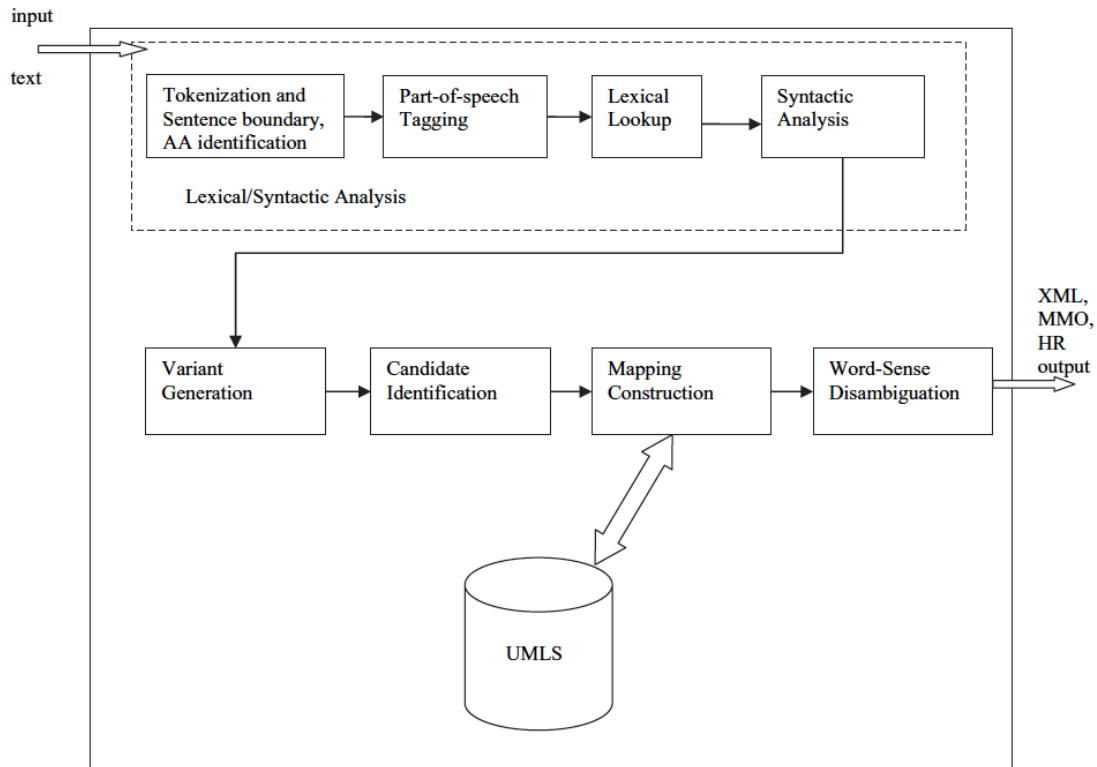
promise to provide controlled domain knowledge to facilitate data integration, information extraction, and decision support was widely recognized. Initially, a number of terminology services were developed that index and maintain ontologies. Some of these engines search the web to find ontologies, such as Swoogle, Watson and OntoSelect, and others provide users with the option to upload their repositories like DAML and SchemaWeb. In 2009, Noy et al from Stanford Center for Biomedical Informatics Research, introduced BioPortal as an open terminology services of biomedical terminologies that accepts different formats and provides automatic updates by user submission and makes those terminologies available through web browsing for human use and Web Service technology for use of applications such as NLP engines (Natalya F Noy et al., 2009). Bioportal is arranged in four logical levels. In the first level, Resources are stored in their original format. The data in the first level is accessed by the second level which is called the Annotation level through format specific access tools, and concept recognition is performed on it using a dictionary of concepts to create a warehouse of Annotation tables. This arrangement results in the abstraction of the format and specifications of the resources from the rest of the system. The information in the annotated tables is then indexed in the third level, which creates an index system to optimize semantic searches received from the fourth level through web browser or web services. Bioportal uses MGrep, a tool developed by the University of Michigan, which implements a novel radix tree based data structure and was therefore found to be faster than UML-Query for matching of text against a set of terminology terms, while implementing the same key idea that is implemented in the mapTold function of UMLS-

Query (N. H. Shah et al., 2009). MGrep is a simple engine and rather than using significant linguistic analysis, relies on a comprehensive lexicon (Aronson & Lang, 2010).

In 2006, UMLS leveraged the MetaMap project, originally developed to improve retrieval of relevant MEDLINE citations based on queries formulated in English, to map the phrases discovered by the SPECIALIST parser to the appropriate concepts in the UMLS Metathesaurus. MetaMap uses an open architecture to perform lexical and syntactic analysis of the input text (Figure 3). This process consists of multiple steps, including tokenization (where sentence boundaries are determined), part-of-speech tagging, lexical look up in SPECIALIST, and a final syntactic analysis to identify the lexical heads in SPECIALIST. Once these lexicons are identified, they go through further processing which includes variants for all phrase words generated and candidates from Metathesaurus are matched to those variations. These matches are then weighted and a ranked list of the best matches is generated. (Aronson & Lang, 2010).

Literature pertaining to evaluation of MetaMap is mostly indirect, in other words comparison of MetaMap's performance with a manual gold standard has almost never been done in a realistic scale (Aronson & Lang, 2010). Most of the studies have involved performing a specific NLP task with and without MetaMap and checking whether the latter improved task performance. The earliest of such studies was performed on MEDLINE articles, as MetaMap was originally developed for MEDLINE. Beyond simple retrieval, studies were conducted to use MetaMap as a medical text indexer.

Figure 3- *MetaMap architecture*



Aronson, R, Lang, FM. *An overview of MetaMap: Historical Perspective and Recent Advances*. *J Am Med Inform Assoc*. 2010;17:229-236

These studies employed document feedback and found that this approach improved the performance, while the final results were comparable with manual indexing (Aronson & Lang, 2010). The BioPortal development group performed a study comparing MetaMap with MGrep in 2009, and found that MetaMap recognized more concepts than MGrep, while the precision and speed of MGrep was better (N. Shah et al., 2009).

Evaluation of NLP in Healthcare Informatics

George Hripcsak and Carol Friedman, from Columbia University in New York, have carried out most of the work in evaluation of NLP systems in the medical domain. In

1997, they published an article describing a set of criteria aimed at improving the quality of NLP evaluation studies and discussed the challenges contributing to the complexity of such task with reference to Message Understanding Conferences (MUC) series of NLP evaluations that were done outside the clinical domain (C. Friedman & Hripcsak, 1998). In this paper, they identified several reasons for the overall lack of evaluations of NLP systems in medicine, including the immaturity of clinical application of NLP, difficulty of evaluation, and lack of published guidelines for evaluating NLP systems in biomedical literature despite such guidelines being present in other domains. This paper focused on comparing an NLP system to some reference standard and not on the impact of such systems on patient care, although the authors note that many of the criteria can also be used in randomized clinical trials that assess such impact. The goal of this article was “to identify measures that will objectively and reliably predict the behavior of the system in a realistic clinical environment” through guidelines that are practical enough that can be followed when possible.

In light of such measures, they reviewed a number of evaluation studies conducted with reference to NLP solutions inside and outside clinical practice. For clinical applications, they considered the evaluation study performed on Linguistic String Project (LSP) system at Glasgow Royal Infirmary. It was found that while the presentation of results was detailed and the description of methods were adequate, the reference standard was weakly described and minimizing of bias was not assured. Another set of studies considered in this section evaluated SPRUS solution, and the authors noted that while in these studies the description of methods and results were found to be satisfactory, there

existed a noted bias of the evaluators being the same as the developers of the system as well as the reference standard of the evaluation. The subject of the next evaluations was CAPIS, which was evaluated in two occasions. In the first instance the study suffered from inadequate analysis of the results and weak reference standard; and the second instance had minimal discussion about the source or type of the captured problems. In the non-clinical domain, multiple MUC studies were evaluated and a number of important findings that had implications for evaluation of NLP system in healthcare were highlighted. The first issue was the aspect of human performance, demonstrating that experts in the reference standard tend to display bias towards the keys that they created for the extraction of concepts, with error rates as high as 30% among themselves. Another important issue was that as the distance between related concepts increased, the accuracy of the detection was reduced. It was therefore noted that for such situations the system required world knowledge and/or linguistic properties (C. Friedman & Hripcsak, 1998). The Friedman and Hripcsak paper established the most prominent challenges of evaluating healthcare NLP solutions. On the top of this list, they note the lack of good performance measures for evaluation of extensibility of a given solution. A system may perform well for one disease in one clinical domain, most diseases in just one clinical domain, a limited number of domains, or all types of clinical data. Most systems perform well in the scope of their design, but fail to keep the same performance when scaled out of their original target. Distinction is required between good performances that are due to NLP system versus one that is the result of the source of data. An experiment in the same paper found an NLP system to have good performance for patient identification when

searched for Parkinson's Disease while not showing as good results when searched for Pneumonia, since the latter could be present as part of patient's past history or some ruled-out differential or suspected diagnosis. Another issue is that there is no way to tie the level of difficulty of the task at hand to the performance measures, since there is no known method that measures the difficulty level of a task in an NLP application. Finally, the strengths and weaknesses of the methodology used by a particular NLP solution have not been studied using real patient documents, and while they are known theoretically, evaluations geared towards these methodologies are not well understood as the prerequisite for such comparison is a common set of clinical documents, a well-specified application, and benchmark measures (C. Friedman & Hripcsak, 1998).

Others have followed in evaluating of NLP solutions in medicine. The Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) was evaluated in an article in 2010, and again one of the most frequent sources of error was the found to be the multilateral relationships of the meaning of an entity to many concepts. Another limitation was found in coordination structure interpretations – For example, 'bowel and bladder habits' was mapped to 'bowel habits' and 'bowel and bladder habits' instead of 'bowel habits' and 'bladder habits'. cTAKES does not have a UML-like Semantic Network functionality that correlates concepts with different semantic types together through predefined relationships. Nevertheless, it is fast, and demonstrated a tokenization accuracy of 95%. it detected the sentence boundaries correctly 95% of the time and attained a part-of-speech tagger accuracy of 93% (Savova et al., 2010). Christensen et al. evaluated the ONYX engine in 2009, and found good inter-annotator agreement of 76-

86% based on the assigned task (least for identifying relationship between concepts, and most for assigning semantic types to relevant words) in the processing of phrases in dental discourse. Many of ONYX's components leverage research in the general and clinical NLP domains, including the use of chart parsing and probabilistic context free grammars. ONYX's use of semantically annotated grammar rules is similar to the semantic grammar approach by MedLEE (Christensen, Harkema, Haug, Irwin, & Chapman, 2009). In 2010, Meystre et al evaluated Textractor, which added machine learning to leverage MetaMap capabilities to extract medications and their prescription justification from clinical narratives. Textractor first analyses the document structure by detection of sections and sentences, and then goes forward to detect tokens and perform part-of-speech tagging. The result is then passed to a module for disambiguation of abbreviations & acronyms and extraction of the drug names and the reason for prescription. The final two stages consist of a context analysis step followed by the extraction of dosage, route, frequency and the duration of treatment. Finally the results were built by joining all the entries for the same medication. Overall precision was found to be 83% for exact matching, and 82-85% (F-measure) for medication information such as dosage or route. However, the accuracy (F-measure) of determining the reason for medication did not reach 28% and correct detection of duration was only observed in 36% of cases. These results further show the diversity in which medical community expresses free form information such as reason or duration of prescription, while the drug names, dosage and route are usually normalized to a finite set of possible values (Meystre, Thibault, Shen, Hurdle, & South, 2010).

Ontologies in healthcare

Ontologies are an explicit definition of terms used in a particular domain. They essentially comprise a model for domain concepts, relationships, properties, and some times, instances of concepts. Ontologies provide a sharable vocabulary in a given domain of discourse for the purpose of common understanding as well as unambiguous information exchange (N.F. Noy & McGuinness, 2001). Examples of ontologies include Amazon product listings, Unified Medical Language System (UMLS) (Bodenreider, 2004), and Systematized Nomenclature of Medicine (SNOMED) reference terminology (Spackman, Campbell, & CÃ, 1997). The distinction between terminologies, or controlled vocabularies, and ontologies sometimes is not well defined (Bodenreider, 2006), especially in the biomedical sciences domain where most of the existing ontologies were started as well organized, but not formally represented, terminologies (W. Ceusters, Smith, & Goldberg, 2005; W. Ceusters, Smith, Kumar, & Dhaen, 2004; Lambrix, Tan, Jakoniene, & Strömbäck, 2007).

Healthcare, compared to other industries, is a unique field in terms of complexity in modeling and knowledge management. Interconnected domains (like administrative, research, clinical) with various degrees of requirements (financial, academic, decision support) leave information management job at healthcare provider organizations level quite a difficult task. Furthermore, ambiguous and context dependent terms and the requirement for multiple levels of granularity of the clinical concepts deepen the complexity of information management at the meaning level. Formal knowledge modeling approaches not only can help rapid extraction of context dependent, consistent,

and comparable information, for the purpose of internal and external interoperability, but also facilitate model driven decision support (Parachoor, Rosow, & Enderle, 2003).

Formal modeling also helps disambiguate terms by making their definition explicit and providing a computational and sharable understanding of heterogeneous information that can be interchanged and easily translated among heterogeneous players (N.F. Noy & McGuinness, 2001). Such models facilitate formalizing concept definitions as well as adding expressivity and reasoning functionalities to the knowledge management system, and thus, improving interoperability among disparate sources of data (Stojanovic et al., 2004). Two other useful advantages of creating and applying a formal model in integrated environments include acquisition of new knowledge and validation of an existing knowledge-based system (Chong et al., 2003).

With regard to the recent trends in information management systems, for moving from silos to more sharable repositories, interoperability and clinical data analytics are gaining momentum within healthcare enterprises (Brailer, 2005; Walker et al., 2005). However, due to the wide range of interconnected domains of care and the existence of multiple players in large healthcare organizations, interoperability still remains as one of the grand challenges (Rossi Mori & Consorti, 1998) where heterogeneous users with heterogeneous data elements and models are involved (Chong et al., 2003; Pisanelli & Gangemi, 2004; Weng et al., 2007). It has been shown in non-healthcare related fields that semantic modeling approaches can be used effectively for interoperability operations among diverse environments (Magoutas, Halaris, & Mentzas, 2007).

Building, extending, and enriching ontologies are achieved in several phases. The first step in ontology development is the selection of a domain of discourse followed by a search in that domain for possible reuse of existing ontologies. In the next step all concepts, terms, and relationships (N.F. Noy & McGuinness, 2001) are identified from various sources. This process can be facilitated by analyzing existing domain documents, through automated and semi-automated concept extraction methods, and consultation with subject matter experts. Also, incorporation of synonymous terms and enrichment of ontologies with additional terms and relationships significantly increases usability of the ontology (Madani, Sittig, & Riben, 2010). In order to build a comprehensive model, all attributes of a given quality metric should be collected and included in the model (inclusive approach). The more comprehensive and enriched the model the more accurate predictions can be made from that model for analytical purposes.

Formal definition of the concepts is the final steps in the ontology development process. Relationship definition among concepts, in the form of concept properties, and application of logical restrictions to the defined concepts will be done in the ontology editing environment. Subsequently, hierarchical relationships are defined and concepts are categorized under corresponding classes (Harris, 2008; N.F. Noy & McGuinness, 2001).

Existing work in clinical information extraction using ontologies

Development and application of ontologies in the domain of quality measurement have been recently become the focus of some researchers. Lee et al.(Lee, Tu, & Das, 2009) evaluated Virtual Medical Record (VMR) (Johnson, Tu, Musen, & Purves, 2001) method

within Standard-Based Sharable Active Guideline Environment (SAGE)(Tu et al., 2007) for the purpose of extraction of cancer quality metrics from EMR systems and concluded that the VMR approach requires additional extensions in order to capture temporal, workflow, and planned procedures concepts. They also emphasized on the fact that patient perspective of care is an important aspect of the overall patient care picture and should be added to the VMR model.

In a short study by Hung (Hung & Stetson, 2007) ontological modeling was evaluated for National Quality Forum's endorsed cardiovascular related quality metrics. The analysis was limited to the evaluation of modeling languages, identification of high-level domain concepts, and percentage of reference terminology coverage for concept components.

Soysal et al (Soysal, Cicekli, & Baykal, 2010) developed and evaluated an ontology-driven system for information extraction from radiology reports. Their objective was to derive an information model from the narrative texts using ontology-driven approach and manually created rules. However, performance-wise, they only evaluated relationships extracted from the narrative texts.

In molecular biology domain Kim et al, showed how a semantic inference module based on domain knowledge can extract regulatory events on gene expression and cell activities. They evaluated extraction results of their system for complex concepts against manually annotated corpora and concluded 53% accuracy. (Kim & Rebholz-Schuhmann, 2011)

There have been other studies that focused on extraction of information, using ontology, from clinical literature. Mildward et al described an interactive method in their study that

enabled end users to refine a given search query from scientific data bases (EMBASE & MEDLINE) and export the results into a structured database. They argued that using a domain specific ontology within semantic queries can potentially enhance response's recall coupled with decent precision (Milward et al., 2005). Muller et al. also looked into mining literature with ontology tools and extract relevant information accordingly. They developed an ontology of domain terms, populated it with instances retrieved from test parsers, and executed keyword-based queries on ontology instances. They argued that using keywords within ontology can increase recall rate from 45% to 95% (Muller, Kenny, & Sternberg, 2004).

National Quality Forum has recently initiated an effort for endorsing and modeling quality metrics with the collaboration of Health Level 7 Standard Community. The new data standard is based on HL7 Reference Information Model (RIM) objects and is called Health Quality Measures Format (HQMF). This standard is related to the HL7 documentation standard (Clinical Documentation Architecture) but is considered as a special, and separate, electronic documentation standard derived from the Clinical Document Architecture ("National Quality Forum Quality Data Model,"). With regard to our previous experience in implementation of HL7 clinical documentation architecture and recent heated debates about shortcomings of HL7 standards (Landgrebe & Smith, 2011) we believe a new approach is needed to fill in the semantic gaps within the proposed standard.

Identification of section headers within clinical narratives is not an easy task. While transcription departments in relatively large hospitals tend to follow a standard for section

headers, healthcare providers are often allowed to create their own version of section headers in clinical notes.

Denny et al (Denny et al., 2009) trained a classifier on a dataset of 10,677 notes based on boundary detection and manual annotation of section headers (95% training, 5% test data). He reported precision and recall of 95.6% and 99% respectively.

In another study by Li et al (Li, Lipsky Gorman, & Elhadad, 2010) Hidden Markov Model (HMM) was used for section header classification within clinical notes. They labeled section with 15 pre-defined section categories (like Past Medical History). The classifier achieved a pre-section and per-note accuracy of 93% and 70% respectively within a dataset of 9,697 clinical notes (78% training, 22% test data).

Tepper et al (Tepper, Capurro, Xia, Vanderwende, & Yetisgen-Yildiz, 2012) described an automatic approach for section segmentation and classification using machine-learning techniques. They calculated a total F-Measure of 92.1% and 90.8% on two different datasets from 374 discharge summaries. They were also able to show the application of section identification for comorbidity extraction from clinical text. Four comorbidities (Diabetes, Hypertension, Asthma, and Sleep Apnea) were targeted for extraction from 14 relevant section headers within 435 discharge summaries of 402 patients. Irrelevant sections such as Family Medical History were eliminated in the final query. The results showed a total of 8% increase (micro-average) in F-Measure for the 4 mentioned comorbidities. Their approach was different from the previous two because it required a small dataset of annotated notes without trusting custom coding for section header boundary identification.

Chapter 3: Methodology

The main objective of this research is to evaluate an ontological approach for extraction of quality of care metrics. Semantic modeling, and in particular an ontological approach, could potentially alleviate the ever increasing problem of information extraction from narrative texts and data integration among interoperable systems in heterogeneous environments like healthcare (Bianchi et al., 2009; Brinkley, Suciu, Detwiler, Gennari, & Rosse, 2006; Burgun, Golbreich, & Jacquelinet, 2004; Ingenerf, Reiner, & Seik, 2001; Pisanelli & Gangemi, 2004).

Hypothesis: can ontological layers enhance the performance of a base NLP output and facilitate unambiguous information extraction from narrative data sources and overcome the current barriers for manual extraction of quality metrics while requiring equal or less time and cost?

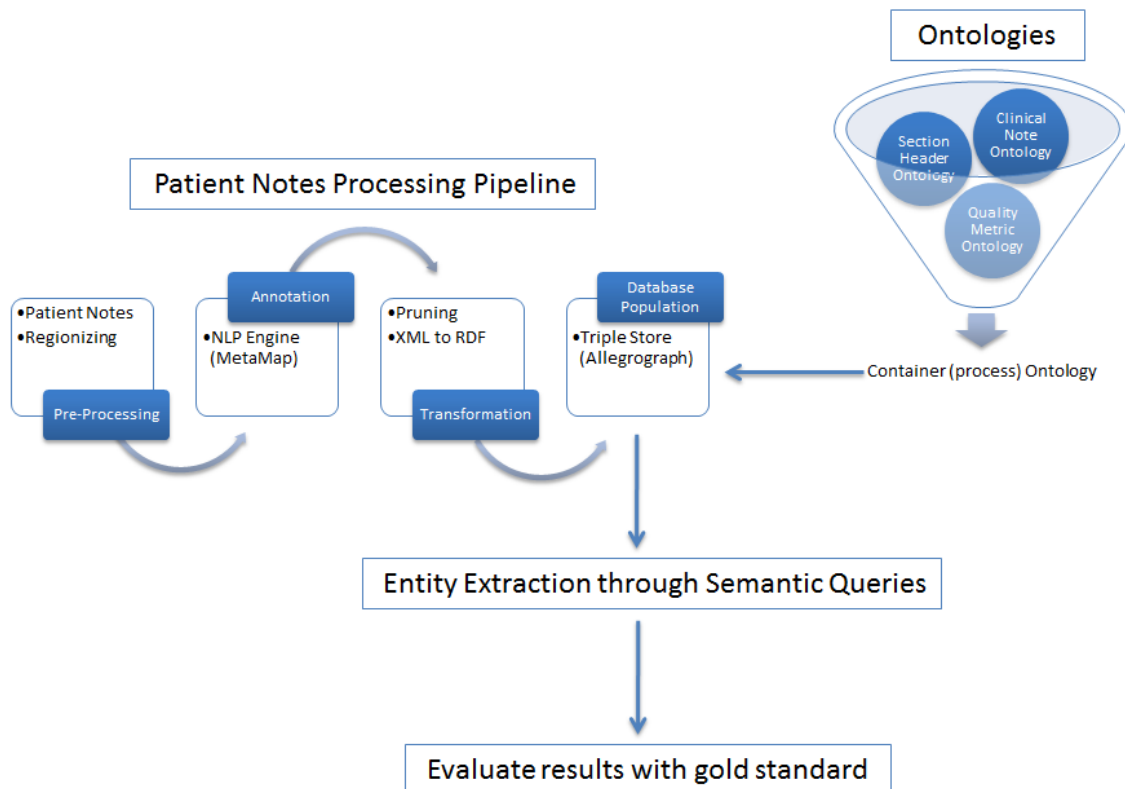
MD Anderson Cancer Center maintains 17 years' worth of narrative (transcribed) documents (>10 million) with 100,000 narrative texts being added every month to its EMR system. If ontology based information extraction proves useful it can then be applied toward millions of clinical notes in all the domains of care at MD Anderson Cancer Center and significantly improves the efficiency of quality metric extraction and reporting process.

Our main contribution is similar to the previous efforts for applying ontological modeling, grounded in the theories of semantics, for the purpose of explicit and

unambiguous extraction of quality metrics but mainly in the domain of healthcare and exclusively in cancer care practice.

The proposed framework for creating an ontology based information extraction of quality metrics could potentially eliminate obstacles in manual metric abstraction from narrative documents. Such complementary addition to existing information extraction system helps enterprise application data integrate more efficiently in terms of data exchange (time & cost) and analytics as part of the enterprise reporting system. A schematic view of our proposed framework is depicted in Figure 3. We explain all the components of this diagram in detail in the upcoming sections.

Figure 4- Schematic view of the proposed ontology-based quality metric extraction framework



Building Ontologies

We have developed three ontologies for our proposed framework with the methods described below:

- a) Identification of the root concept in SNOMED CT and concept hierarchy for each of the 5 selected quality metrics in our study (e.g., Diabetes). We used View Extraction functionality ("View Extraction - NCBO Wiki," 2013) within National Center for Biomedical Ontology (NCBO) Bioportal Appliance ("Welcome to the NCBO BioPortal | NCBO BioPortal," 2013) in order to traverse and extract all children of a given parent concept. We implemented a local version of NCBO appliance for faster response time.
- b) Binding standard terminological codes (SNOMED CT and RxNorm) to the components derived in the previous step and creating concept ontology.
- c) Processing clinical narrative documents in the target patient group and extracting section headers (like Past Family Medical History) from them.
- d) Building a section ontology from extracted section headers in the previous step
- e) Building a clinical note ontology that contains patient note meta data, relevant section headers, and the relevant concepts

Ontology Language & Editing Environment

The standard language for semantic web is Resource Description Framework (RDF). Other standard semantic web languages, with higher degrees of expressivity, such as Ontology Web Language (OWL) (McGuinness & Van Harmelen, 2004a), have been used for semantic web modeling or ontology engineering (Allemang & Hendler, 2008; N.F. Noy & McGuinness, 2001). Simple Knowledge Organization System (SKOS) ("SKOS

Simple Knowledge Organization System - home page," 2013) is also introduced by Semantic Web working group as a new standard for knowledge organization such as thesauri and classification schemas. For the purpose of our modeling and in order to establish broader interoperability we have selected the OWL standard for patient note and concept ontologies and SKOS for section header ontology ("SKOS Simple Knowledge Organization System - home page," 2013). SKOS standard includes properties such as “*narrower than*”, “*broader than*”, and “*exactMatch*” which make it more suitable for our section header classification purposes.

TopBraid Composer™ ("TopQuadrant | Products | TopBraid Composer,") and Protégé (N.F. Noy & McGuinness, 2001) are used as the ontology editing environment for semantic modeling in OWL format. The build process will include formal definitions of concepts and their relationships as well as terminological bindings to standard reference vocabularies.

Patient Selection

Since the focus of our study is ACS NSQIP quality metrics, we have adopted NSQIP guidelines for patient selection. This is a two-step process with systematic sampling dates (when) and inclusion/exclusion criteria (what).

The minimum number of the collected cases is determined by the volume of the surgical cases during a one year period (which is around 15% of total case volume). The sampling process includes the “8-day cycle” selection method which guarantees that all cases have an equal chance of being selected from each day of the week. This is a mandatory case

selection process for assuring that proper systematic sampling is performed on the surgical caseload during a calendar year period.

Based on the contract of the participating hospital with NSQIP and the type of the surgeries performed inclusion and exclusion criteria are applied to the pool of patients identified in the first step of the sampling phase. Operative logs, CPT codes, patient age, in operation room time, and operating room data are analyzed during inclusion/exclusion phase and the final number of the cases determined by site adjusted case requirement calculations. For our study, we included all 2085 patients at MDA that were selected with this method during 2011 calendar year for ACS NSQIP quality metric extraction project. More information about patient selection method is available online from ACS NSQIP website ("About ACS NSQIP | ACS NSQIP," 2013)

Gold Standard

ACS NSQIP guidelines states that metric collections should be done from patient notes and not from administrative, billing, or insurance data. Steinberg et al. showed in their study that administrative and billing data have less consistency and reliability compared to patient charts (notes) for reporting ACS NSQIP quality metrics related to complications and surgical site infection (Steinberg, Popa, Michalek, Bethel, & Ellison, 2008).

In ACS NSQIP program, each hospital has assigned Surgical Clinical Reviewers (SCR) (or abstractors) for collecting quality metrics from various data sources within EMR system. SCRs at MDA spend an average of 45-60 minutes per patient to abstract ACS NSQIP quality metrics. Required quality metrics for collection and reporting to ACS

NSQIP are categorized in different groups which are documented in the guideline for SCR ("About ACS NSQIP || ACS NSQIPACS NSQIP," 2013). These groups include: Demographics, Surgical Profile, **Preoperative Risk Assessment**, Perioperative Laboratory Data, Occurrences, Postoperative Laboratory Data, Postoperative Information, and Follow ups. Upon interviewing abstractors at MDA Quality Engineering Department, who were responsible for abstraction of ACS NSQIP quality metrics, we found that metrics related to Preoperative Risk Assessment group is the most time consuming part of the abstraction process. These metrics are generally documented in transcribed clinical documents (dictated patient notes) and abstractors have to read such notes in order to report the required quality metrics to ACS NSQIP. It should be mentioned that SCRs are nursing staff who have extensive training in NSQIP abstraction protocols & guideline. They are also actively participating in NSQIP certification, audition, and training programs. Shiloach et al. (Shiloach et al., 2010) looked into inter-rater reliability metric and found 1.56% disagreement rate among SCRs of the participating hospitals in ACS NSQIP program. NSQIP data also shows that reliability has been improved with continuous training and auditing since the start of the program in 2005.

The dataset that we received from MDA Quality Engineering Department included NSQIP abstracted information over a period of 12 months from 2,085 patients who had undergone surgery at MDA in 2011. We've considered this reported operational dataset as the gold standard for our study

Metric selection

Abstractors at MD Anderson report Preoperative Risk Assessment quality metrics as Boolean values (Yes/No) to the ACS NSQIP program. We have selected 5 of these metrics that have a frequency of more than 30 positive cases (Boolean value="Yes") among our gold standard. These metrics include: Diabetes Mellitus, Hypertension, history of Cardiac Surgery, history of CNS tumors, and Transient Ischemic Attack (TIA). The frequency of Boolean values for these metrics is shown in table 2. The complete list of quality metrics and their definition is available from NSQIP website ("About ACS NSQIP || ACS NSQIPACS NSQIP," 2013) .

Table 2 – *Selected NSQIP quality metrics reported values*

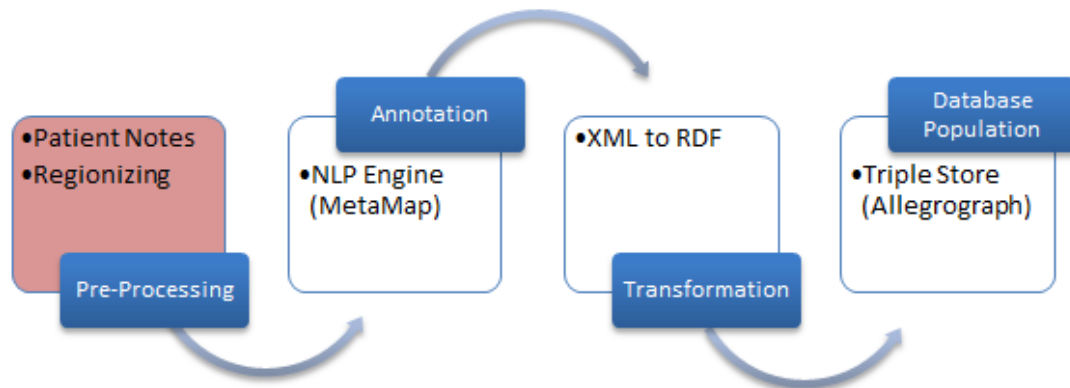
Quality metric	Yes	No
Diabetes Mellitus	227	1,859
Hypertension	906	1,180
History Cardiac Surgery	69	2,017
History CNS Tumors	127	1,959
Transient Ischemic Attach	34	2,052
Total	1,363	9,067

Patient note extraction & pre-processing

All transcribed documents of 2085 selected patients in our study were extracted from MDA EMR repository (Figure 5). Python scripting was used to eliminate unwanted characters and extract section headers. A typical patient note composed of regions of texts. Each region consists of a section header (Chief Complaint, History of Present Illness, Physical Exam, etc.) and its relevant content. Transcriptionists at MDA follow a

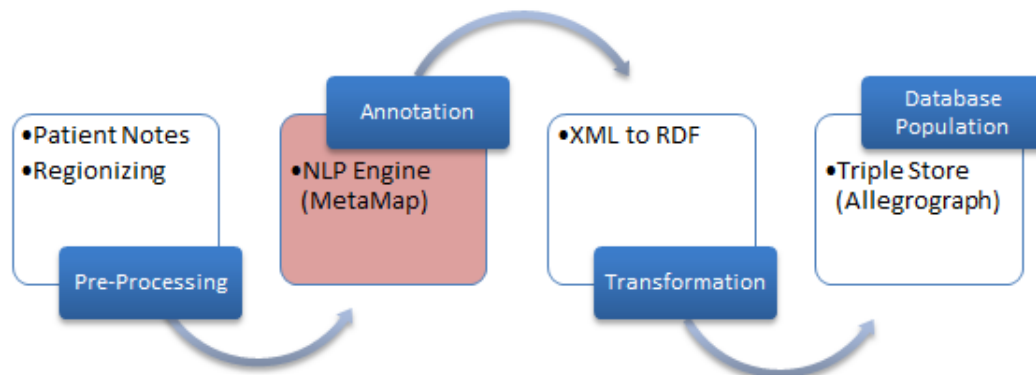
standard guideline to convert dictated voice recordings into narrative texts; Section headers usually start at the beginning of a new line, all in upper case, and end in colon.

Figure 5 – *Patient notes processing pipeline: Extraction and pre-processing*



Using Python scripting, we identified each region of the text within a note, extracted associated section and its content, and converted the data into XML format. We incorporated all extracted section headers into our section ontology built process, as we explained in Building Ontologies section, and sent the resultant XML output to MetaMap for further content analysis (Figure 6).

Figure 6 - *Patient notes processing pipeline: Conversion of processed notes to XML*



Natural Language Processing (NLP) Engine

We implemented National Institute of Health (NIH) natural language processing engine (MetaMap v2012) in order to parse and annotate narrative notes obtained from the pre-processing phase. This application is available free of charge for research community. MetaMap was installed on a Linux 64 bit server behind MD Anderson secure firewall. A custom Python script was written to pull data from processed notes and submit the content of each section header to MetaMap for parsing NLP analysis (Figure 6). After extensive testing and collaboration with MetaMap development team at NIH and in order to reduce the noise in the output we selected below configurable options in MetaMap:

- Word Sense Disambiguation set to active
- Composite Phrases set to active with maximum of 4 prepositional phrases allowed
- Threshold (evaluation score cut off) set to 580
- Terminology was limited to SNOMED CT and RxNorm
- Allowed semantic types were restricted to : Disease or Syndrome , Sign or Symptom, Mental Process, Mental or Behavioral Dysfunction, Acquired Abnormality, Anatomic Abnormality, Diagnostic Procedure, Therapeutic or Preventive Procedure, Neoplastic Process, Finding, Pathologic Function, Congenital Abnormality, Pharmacologic Substance
- XML output format set to XMLf1

One XML file was generated for each patient note (46,835 XML total) that contained additional metadata such as patient id (encrypted), note type, note date, and note service.

Figure 7 below represents the structure of the resulting XML of one sample note. The root node contains patient id followed by the note metadata, regions of texts, and MetaMap put (MMO). The content of each section header was captured in “Body” element of the XML.

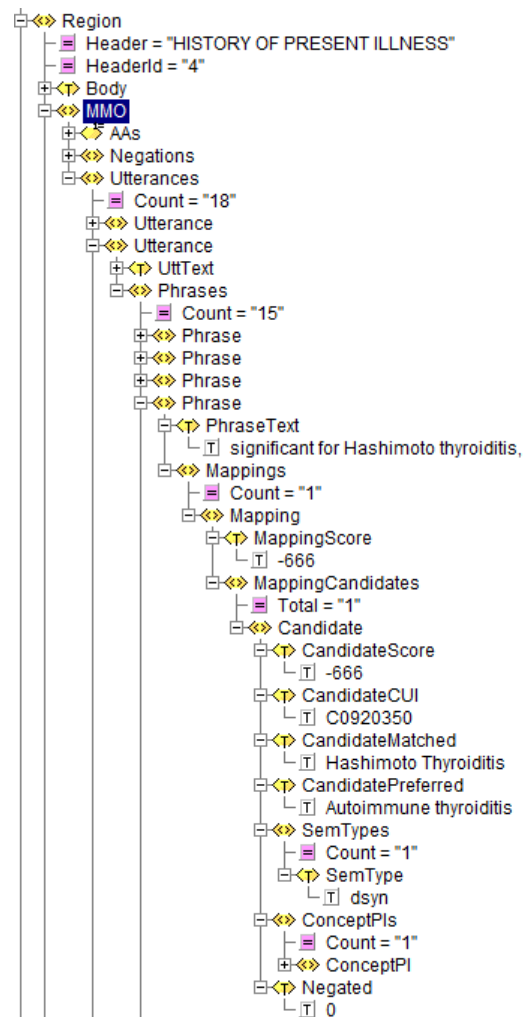
Figure 7 - Parsed and regionized patient note are converted into XML format

The screenshot displays an XML viewer interface. On the left, a tree view shows the hierarchy: Patient (patientId = "d8715e88757ed01302793327693cb9dd") -> Note (date = "7/5/2011", recordId = "41", service = "791", type = "History and Physical") -> Region (Header = "REASON FOR VISIT", HeaderId = "60") -> Body -> MMO. The right pane shows the XML code snippet: <?xml version="1.0" encoding="UTF-8"?><Patient p
<Note date="7/5/2011" recordId="41" service="791"
<AAs Count="0"/>
<Negations Count="0"/>. Below the code is a Tree Selection Browser showing 4 Attributes: date (7/5/2011), recordId (41), service (791), and type (History and Physical). It also lists 16 Subtags: Region (REASON FOR VISIT, HISTORY OF PRESENT ILLNESS, REVIEW OF SYSTEMS, PAST MEDICAL HISTORY, PAST SURGICAL HISTORY, FAMILY HISTORY, SOCIAL HISTORY, ALLERGIES, MEDICATIONS, PHYSICAL EXAMINATION, VITAL SIGNS, GENERAL, ABDOMEN, LABORATORY DATA, IMAGING STUDIES, ASSESSMENT AND PLAN).

Note. The content of each section (Body element) is sent to MetaMap for analysis. The results is captured under the MMO XML element

A sample expanded MetaMap (MMO) node is shown in Figure 8. Each sentence is parsed into phrases. Phrases are analyzed subsequently and mapped to categorical concepts (SNOMED CT & RxNorm). A specific element down in the XML branch shows whether the concept is negated. This is a new extension to MetaMap output based on our request to the MetaMap developer team and will be included in the upcoming MetaMap release.

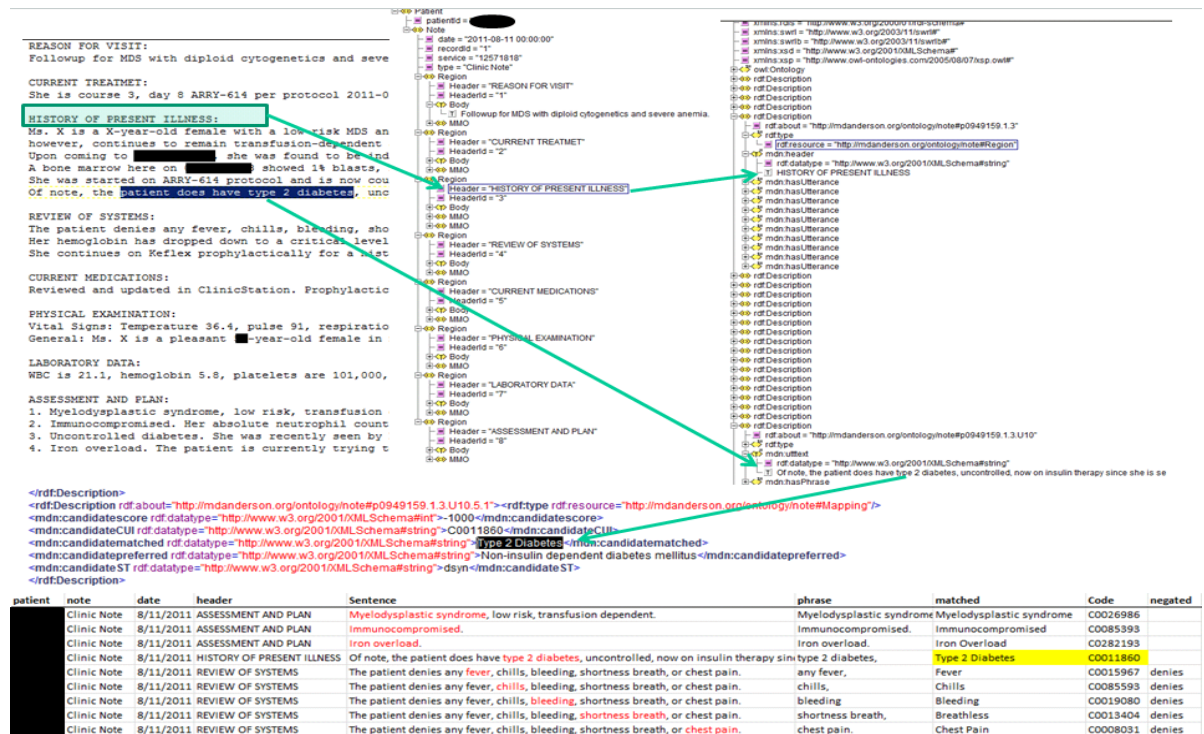
Figure 8 - Expanded view of MetaMap node (MMO) in the XML output



Data format and repository type

In order to decrease the size of the XML data obtained from the previous phase we pruned unwanted XML elements from MetaMap output with custom Python scripting (like start and stop positions, etc.). Subsequently we converted XML data into RDF format and loaded them into a RDF repository. We selected AllegroGraph® repository ("AllegroGraph RDFStore Web 3.0's Database," 2013) since MDA has a purchased license and support for it. We also used SPARQL Protocol and RDF Query Language ("SPARQL Query Language for RDF," 2013) to perform federated query across different ontologies and the RDF repository. The complete processing pipeline and query results for a sample note and quality metric (Diabetes) is shown in Figure 9.

Figure 9 - Conversion of a processed note into RDF and extraction of a quality metric



Evaluation of Ontologies

In order to use a domain specific ontology in an information extraction system, its content, structure, and function should be validated to ensure all requirements for maximum content coverage, consistency, and usability are met.

Many frameworks have been proposed for evaluation and validation of ontologies in the biomedical realm (Brank, Grobelnik, & Mladenic, 2005a, 2005b). The evaluation of an ontology usually consists of verification and validation processes. Logical rule engines are used for verification of logical, terminological, and structural consistencies. Subject matter experts are being consulted for domain coverage and completeness of the ontology (Obrst, Ashpole, Ceusters, Mani, & Smith, 2007). For complex reasoning, rule (or inference) engines are less used in the biomedical field compared to other computational fields; however, despite existing discrepancies in the structure of some of the clinical ontologies (like SNOMED CT) (Carlson et al., 2010; W. Ceusters, Smith, & Flanagan, 2003) logical reasoners have been used for validation of the classification (Wolstencroft, McEntire, Stevens, Taberner, & Brass, 2005) and part-whole analysis (Hahn & Schulz) as well as verification of structural integrity of the ontology. Other methods for ontology validation in the field of biomedicine include: application usage, data source coverage, benchmarking against an existing ontology, and criteria based assessment (Obrst et al., 2007) .

We have chosen formal methods and statistical agreement tests for evaluating structure, domain coverage, and function of our ontological framework. Formal methods are compatible with the requirements identified by the World Wide Web Consortium (W3C)

Provenance Incubator Group ("W3C Provenance Incubator Group Wiki,"), evaluation framework for controlled medical vocabularies (Cimino, 1998), ontology of diseases (Bodenreider & Burgun, 2009), and the Open Biological and Biomedical Ontology (OBO) Foundry (Smith et al., 2007). Statistical agreement tests validate application usage and criteria based assessments of the model (functionality) against available data from manual abstraction (gold standard).

I - Formal Methods

Content Coverage

1. Existence of mapping to clinical terminologies: we aligned our model with selected UMLS Metathesaurus (RxNorm & SNOMED CT) so that the optimal terminological bindings are acquired.
2. We consulted domain experts (abstractors) for inclusion and exclusion of concepts derived from the ontology building phase for maximum content coverage and relevancy.

Structure

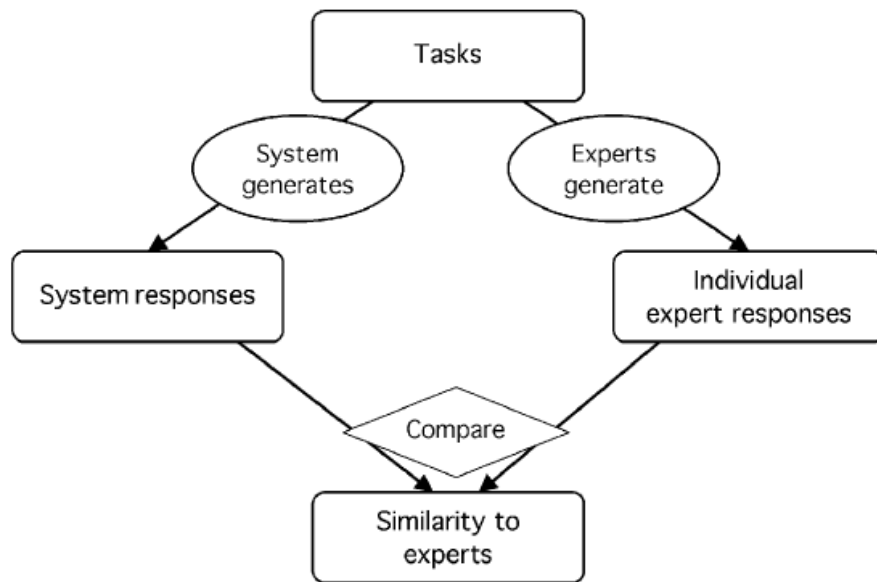
1. Dynamic classification with existing rule engines was used within our ontology editing environment (TopBraid Composer™) for structural validation of the model. A well- structured model should not generate any error during verification process by the rule engine.
2. Provider friendliness and standard representation format of the derived model was evaluated by subject matter experts. The current ontology format corresponds to OBO

Foundry principal number two ("Open Biological and Biomedical Ontologies: Archive of original principles," 2013; Smith et al., 2007).

II – Statistical Agreement Tests

Many evaluation techniques have been defined for health informatics applications in different areas such as system, outcome, impact, and cost effectiveness. For system evaluation in information extraction tasks usually a computer output is compared with a gold standard (human) output (George Hripcsak & Rothschild, 2005). Our goal of evaluation is to quantify how much a system performs like an expert. In our study, our subject matter experts (abstractor or clinical trained nursing staff) generated a reference standard as part of their operational data reporting activity. According to Hripcsak et al (George Hripcsak & Wilcox, 2002) different models can be used for evaluation of information extraction system in which subject matter experts can play different roles. If SMEs are tasked to quantify performance of an information extraction system, they can either generate a reference standard (abstraction of clinical notes in our case) or judge the output of a system generated output. SMEs can also play the role of comparison subjects for interpretation of a comparison study with an information extraction system. In our study we consider SME generated data as a gold standard and compared our ontology based information extraction system results with the gold standard (Figure 10) (George Hripcsak & Wilcox, 2002).

Figure 10 - Use of subject matter expert for comparison of the data generated by a system



Hripcsak G, Wilcox, A. Reference Standards, Judges, and Comparison Subjects: Roles for Experts in Evaluating System Performance. J Am Med Inform Assoc. 2002;9:1-15.

In order to calculate the agreement rate between our ontology based information extraction and the manual abstraction method (gold standard) we used precision, recall, and F-Measure metrics.

In information retrieval methods, two primary metrics have been suggested for quantification of agreement between two responses; Precision and Recall (George Hripcsak & Rothschild, 2005). Precision, that is also called positive predictive value (PPV), is the ratio of the number of relevant findings retrieved to the total number of findings retrieved. Using a contingency table, precision is calculated by dividing true positives by true plus false positives (Table 3 & Equation 1). On the other hand, recall is the ratio of the number of relevant findings returned to the total number of the findings. Recall metrics is similar to sensitivity of a system and can be calculated by dividing true

positives by true positives plus false negatives (Equation 2). The agreement between any two sets of responses from information retrieval systems can be calculated by these two metrics.

In order to obtain a harmonic balance between precision and recall they are often combined and presented as F-Measure which is simply a calculated balanced value of these two metrics (Equation 3).

Table 3 – *Contingency table*

		Reference Standard		
		Yes	No	
IE System	Yes	TP	FP	(PPV)
	No	FN	TN	(NPV)
		(Sens)	(Spec)	

Note. True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN)

Equation 1

$$Precision \text{ (Positive Predictive Value)} = \frac{TP}{(TP + FP)}$$

Equation 2

$$Recall \text{ (Sensitivity)} = \frac{TP}{(TP + FN)}$$

Equation 3

$$F = \frac{(1 + \beta^2) \times recall \times precision}{(\beta^2 \times precision) + recall}$$

In most evaluations, $\beta = 1$ in F Measure equation but if in special use cases where false positives or false negatives have considerable implications, therefore weighed heavily, a different value can be assigned to β for a more tailored value of F-Measure (George Hripcsak & Rothschild, 2005). In our study we assumed a value of 1 for β . The higher value of F-Measures shows a higher agreement between two systems.

Two other metrics have been proposed for agreement studies between 2 or more systems (George Hripcsak & Rothschild, 2005); Agreement and Agreement beyond chance (kappa test). Equation 4 shows how to calculate simple agreement between two raters. It is simply the proportion of occurrences where the two rating systems agree.

Equation 4

$$Agreement = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

However, if TN counts are large (like our case) this formula masks positive cases values and causes the equation to lean toward 1. For such situations, where the number of true positive cases is small relative to the true negative cases, positive specific agreement is used and can be calculated by equation 5

Equation 5

$$P_{Pos} = \frac{2 \times TP}{(2TP + FP + FN)}$$

This formula is similar to the F-Measure formula shown in equation 3 when $\beta=1$.

Agreement beyond chance (kappa) can also be calculated as shown in equation 6

Equation 6

$$k = \frac{2 (TP \times TN - FP \times FN)}{(TP + FP)(FP + TN) + (FP + TN)(TP + FP)}$$

However, in cases where TN numbers are unknown or high (like in our case) kappa leans toward F Measure again (Fleiss, 1975; George Hripcsak & Rothschild, 2005). Therefore, we will be using precision, recall, and F Measure metrics for reporting our agreement results between the two systems for each of the extracted quality metrics.

In order to aggregate the results of agreement measurements (Precision, Recall, and F-Measure) from all extracted quality metric we used two methods for averaging the results; Micro-Averaging and Macro-Averaging. When there are multiple classes of contingency tables, averaging the evaluation scores provides a more general picture of all class results (Van Asch, 2013). Micro-averaging is the most common averaging method in which each *extracted instance* is given the same weight. Because TN is not included in F Measure calculation the score is largely determined by TP cases, hence, quality metrics with large number of TP dominates micro-average. Micro-average is calculated from the aggregated values that are *pooled* from each contingency table into a target pooled table (Table 4).

Table 4 - *Micro-averaging multiple contingency tables*

Metric 1			Metric 2			Pooled		
	Y	N		Y	N		Y	N
Y	20	10	Y	80	10	Y	100	20
N	10	160	N	10	100	N	20	260

$$Precision = \frac{100}{100+20} = 0.83$$

In the second method (or Macro-Averaging) each metric is given the same weight but averaging is done by a traditional averaging method; combining calculated agreement values (Precision, Recall, F Measure) from each contingency table and dividing them by the number of contingency tables (Table 5).

Table 5 - *Macro-averaging multiple contingency tables*

Metric 1			Metric 2		
	Y	N		Y	N
Y	20	10	Y	80	10
N	10	160	N	10	100

$$Precision\ 1 = \frac{20}{20+10} = 0.67 \quad Precision\ 2 = \frac{80}{80+10} = 0.89$$

$$precision = \frac{0.67+0.89}{2} = 0.78$$

Chapter 4: Findings

Transcribed documents

Originally, 191,645 dictated notes associated with our patient population (2,085) were extracted from MD Anderson EMR repository. These notes are categorized under 3 major groups in the EMR system: Radiology, Pathology, and Transcribed documents. 60,808 notes were identified as Pathology and Radiology notes in our dataset.

Table 6 - *Top 20 Transcribed Patient Notes types and their frequencies*

Note type	Frequency
Clinic Note	45,478
Progress Note	14,737
Consultation	12,731
Telephone Note	9,617
Operative Report	7,832
XRT Clinic Note	5,094
History and Physical	5,070
Discharge Summary	3,304
Nutrition Follow Up Note	2,603
Social Work	2,516
Procedure Note	2,194
Nursing Note	1,953
Nutrition Assessment Note	1,942
Primary Medical Evaluation	1,825
Brief Operative Procedure Note	1,581
Study Entrance Note	1,403
Day of Proc History and Physical Update	1,199
XRT Simulation Note	1,189
Emergency Room Note	1,060

Transcribed documents (the remaining 130,837 notes) included 48 note types such as History & Physical, Social work, Consultation, etc. The top 20 most frequent transcribed document types are shown in Table 6.

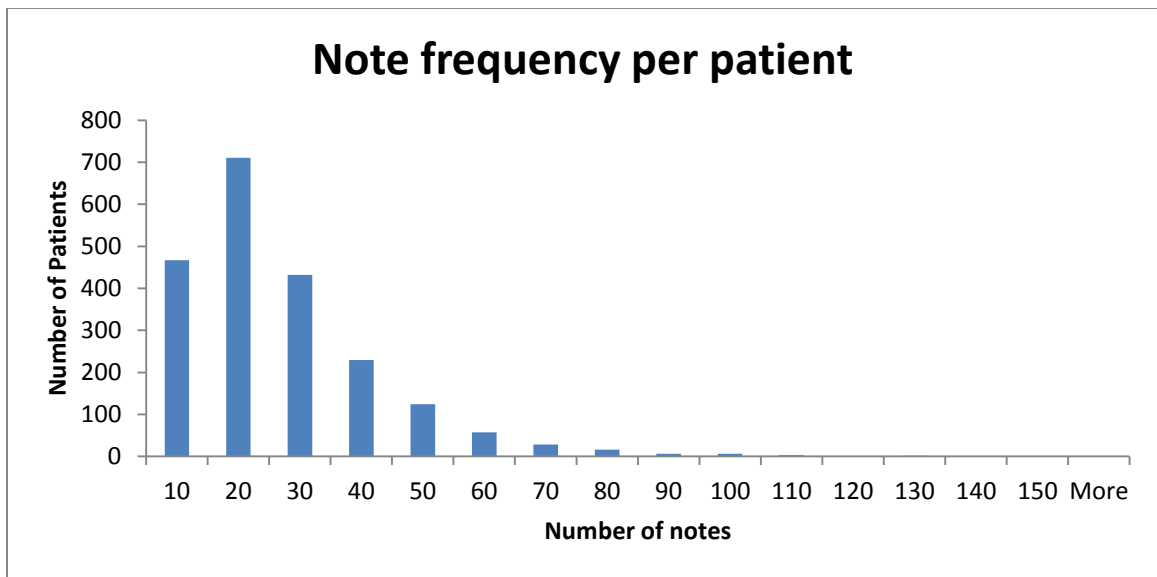
According to abstractor's guideline only 8 transcribed note types were reviewed in the manual abstraction process during 2011. To be compatible with the gold standard data, in terms of the note types and note dates, 144,810 notes were excluded from our study.

Table 7 shows frequency of the 8 selected note types, number of the section headers found in these notes, and the average number of section headers extracted per note for each note type. Within this filtered pool of 46,835 patient notes of our study, the highest and lowest number of notes per patient was 148 and 1 note(s) respectively with the average of 22 notes. The distribution of the number of notes per patient is shown in Figure 12

Table 7 - *Selected 8 note type frequencies and section header counts*

Note type	Count	Section header	Average section header per note
Clinic Note	20,491	180,378	8.8
Consultation	7,808	110,983	14.2
Operative Report	5,686	62,590	11.0
Telephone Note	5,367	11,579	2.2
History and Physical	3,107	53,382	17.2
Discharge Summary	2,201	29,547	13.4
Procedure Note	1,094	8,496	7.8
Primary Medical Evaluation	1,081	18,782	17.4
Total	46,835	475,737	11.5

Figure 11 - *Distribution of note counts per patient*



Ontologies

Section Header Ontology

Extracted section headers from patient notes, that were extracted from EMR and pre-processed, were used in building the section header ontology.

In order to evaluate our section header extraction algorithm we randomly selected 500 notes (100 noted from each identified quality metrics category) and evaluated for precision and recall. Notes were examined by subject matter experts, annotated for section headers, and compared with our automated section header extraction algorithm. Results are shown in a contingency table (Table 8) where the number of true positives, false positives, and false negatives are captured and used in calculating precision, recall, and F Measure.

Table 8 - *Automatic section extraction performance compared to a gold standard*

System	Gold		
		Yes	No
	Yes	8391	90
	No	242	
	Precision	Recall	F-Measure
	0.99	0.97	0.98

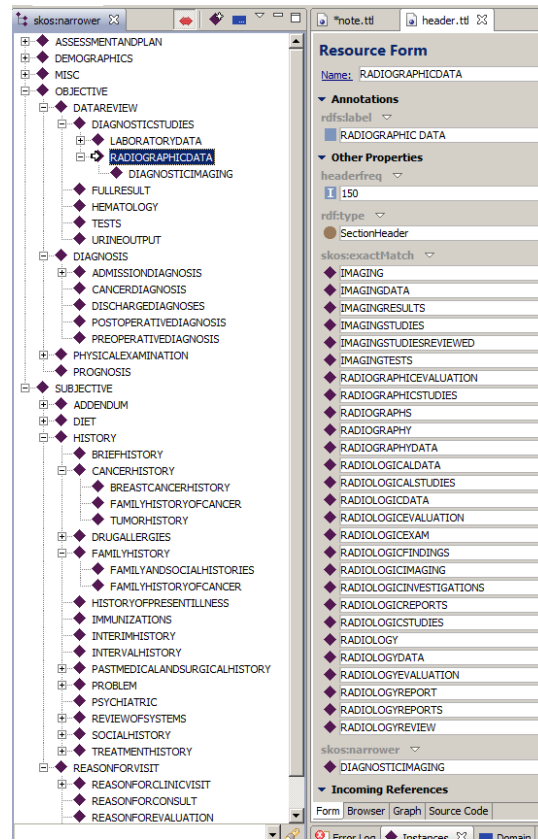
We used SKOS *narrower than/broader than* and *exactMatch* properties ("SKOS Simple Knowledge Organization System - home page," 2013) for classifying section headers into hierarchies and assigning synonyms respectively (Figure 12).

Each section header is examined and categorized as relevant (to be included in the query) or irrelevant (not included in the query) after getting feedbacks from subject matter experts. The distribution of section headers for each metric is shown in Table 9. Relevant section examples include Assessment, Medical History, and Impression. Irrelevant section examples include Family Medical History, Recommendation, and Complications.

Table 9 - *Section header distribution within 5 selected quality metrics*

	Unique section header count	Relevant	Irrelevant
Cardiac Surgery	104	64	40
CNS Tumor	257	175	82
Diabetes	224	122	102
TIA	51	39	12
Hypertension	279	174	105
Total	915	574	341

Figure 12 - Section header classification and synonym assignment using SKOS



Quality Metric Ontology

We identified the root concept for each selected quality metrics in SNOMED terminology (Jan 2013 version) and extracted all of their relative children. The SNOMED root concepts include:

- Operation on heart (Cardiac surgery procedure), ID 64915003
- Neoplasm of Nervous System (Tumor of nervous system), ID 126950007
- Diabetes Mellitus (DM), ID 73211009
- Hypertensive disorders (Hypertension), ID 38341003
- Transient cerebral ischemia (TIA) , ID 266257000

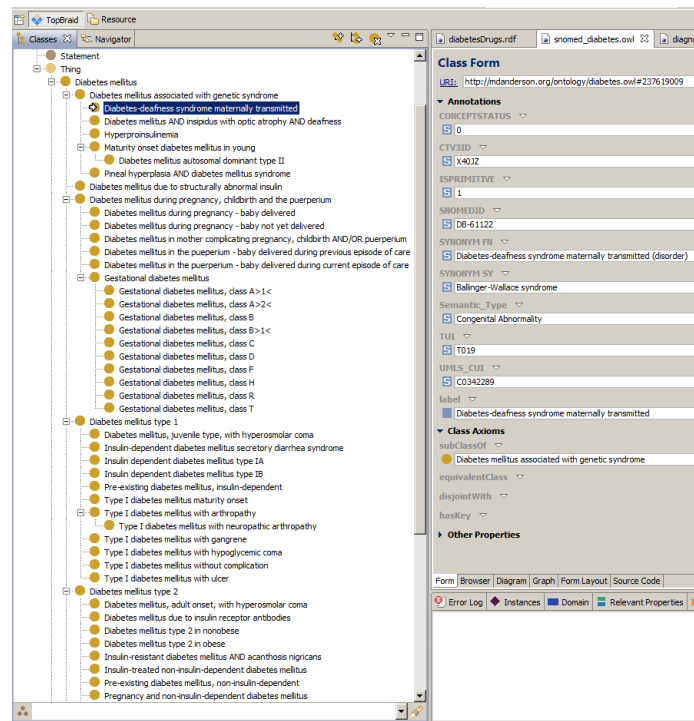
Concept count and a sample concept ontology is shown in Table X and Figure X

Table 10 – *Number of concepts included in quality metric ontology*

Ontology	Concept count
Diabetes	91
Cardiac Surgery	958
Hypertension	106
CNS Tumors	835
TIA	11

According to the metric definition for diabetes Mellitus, patient should also take a diabetes related medication in order to be reported as a diabetic patient. For this purpose, we have also created an ontology of diabetes mellitus medications, with mappings to RxNorm, from the same reference that abstractors used to match patient medication with diabetes ("Patient Handout - Diabetes Medicaiton," 2013) (Appendix A).

Figure 13 - *Diabetes Mellitus ontology hierarchy*

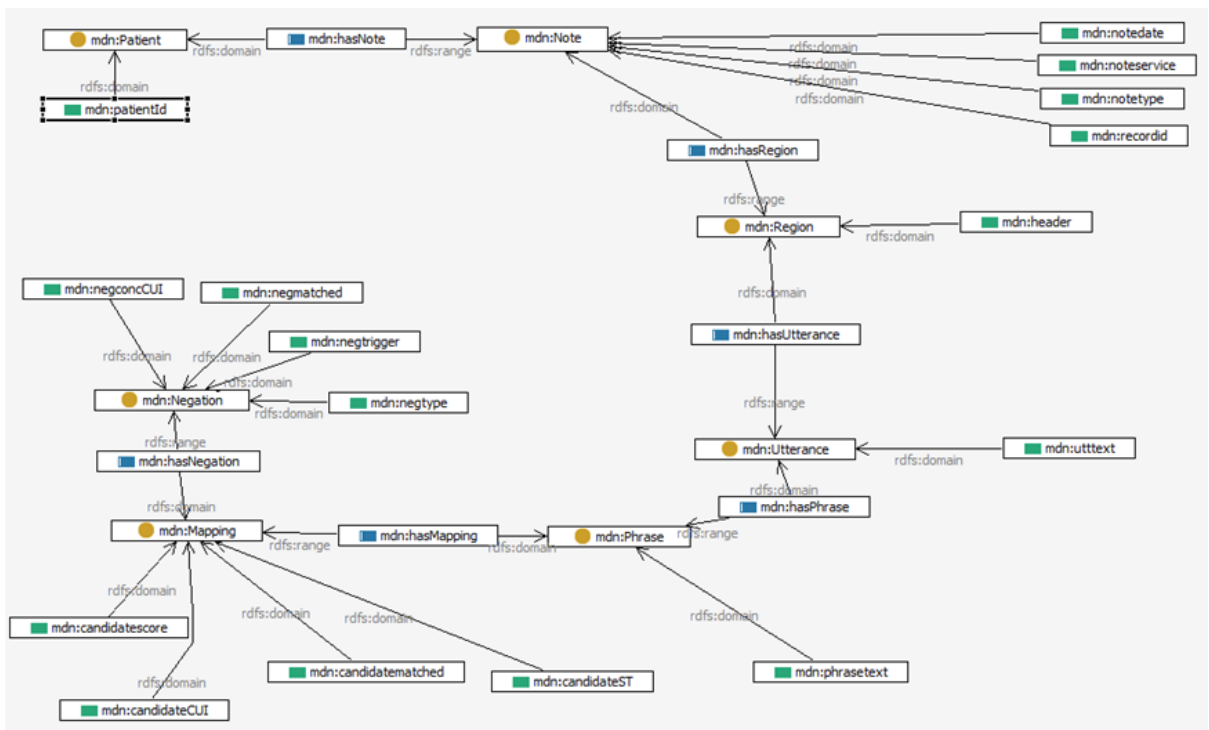


We also reviewed concept ontology with abstractors and eliminated irrelevant concepts. For example, concepts like *Maternal diabetes mellitus*, *Gestational diabetes mellitus*, *Maternal hypertension*, *Pre-eclampsia*, *Renal sclerosis with hypertension*, and *Diastolic hypertension* were excluded from the concept ontology.

Clinical Note Ontology

For this ontology we created seven main classes and build the relationship among them; Patient, Note, Region, Utterance, Phrase, Mapping, and Negation classes. The relationship between these classes and associated properties are shown in Figure 14 below.

Figure 14 - Patient note ontology: Objects are shown in gold, objects properties in blue, and data type properties in green



We populated all the RDF instances (46,835), described in the method section, into the patient note ontology within AllegroGraph repository. Number of the instance counts and associated data type properties for each class is shown in Table 11

Table 11 - *Instance count of the main patient note ontology objects*

Object	Instance count	Object Metadata
Patient	2,085	Patient id
Note	46,835	Note type, Note date, Note service, Note id
Region	475,691	Section header
Utterance	2,343,856	Utterance text
Phrase	11,627,224	Phrase text
Mapping	3,263,338	Semantic Type, Mapped SNOMED concept, Mapped CUI, Score
Negation	535,205	Negation trigger, Negation type, Negated Concept, Concept CUI
Total	18,294,234	

Our repository contained 70,907,728 triples. The difference between instance count and triple count shows the number of relationships that exist among instances of classes shown in Figure 14. Simple SPARQL queries within populated patient note ontology effortlessly pinpoints identified concepts (with mappings to SNOMED) under associated phrase, sentence, section header, and patient note. We will use this structured format for filtering unwanted concept (from concept ontology), non-negated concepts, and irrelevant sections (from section ontology) in our federated queries. In the next section we'll discuss how adding multiple layers of ontology (section ontology, concept ontology) and context (negation) can affect the precision, recall, and F-measure of the base NLP output.

Figure 15 - Sample query from instances populated in the patient note ontology

[patient]	[note]	[date]	[header]	[utt]	[phrase]	[matched]	[cu]	[st]	[negtrig]
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	PLAN	The patient should follow up in 1 year or sooner if there are any changes.	follow up in 1 year	Follow-up 1 year	C0420338	findg	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	PLAN	I have recommended AREDS supplementation and using the Amsler grid.	using	used	C1273517	findg	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	PLAN	I have recommended AREDS supplementation and using the Amsler grid.	recommended AREDS su...	Supplementation	C0242297	topp	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	DEPRESSION		Macular degenera...	Macular Degeneration	C0242383	dym	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	DEPRESSION	Dry macular degeneration, level 2 to 3.	macular degeneration,	Macular Degeneration	C0242383	dym	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	PHYSICAL EXAMINATION	The disc, vessels, and periphery are otherwise normal.	vessels,	VAS	C0042815	dap	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	PHYSICAL EXAMINATION	There is no subretinal fluid or hemorrhage.	no subretinal fluid	Subretinal fluid	C1720732	findg	no
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	PHYSICAL EXAMINATION	Dilated fundus examination shows semi-confluent drusen in the per-fov...	semi-confluent drusen in ...	Confluent drusen	C1720180	dym	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	PHYSICAL EXAMINATION	Slit-lamp examination is only remarkable for 1+ nuclear sclerosis, both e...	eyes.	Ophthalmia	C0014236	dym	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	PHYSICAL EXAMINATION	Slit-lamp examination is only remarkable for 1+ nuclear sclerosis, both e...	only remarkable for 1+ n...	Nuclear sclerosis	C0392557	dym	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	PHYSICAL EXAMINATION	Slit-lamp examination is only remarkable for 1+ nuclear sclerosis, both e...	Slit-lamp examination	Slit lamp examination	C0419360	dap	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	PHYSICAL EXAMINATION	On examination visual acuity is 20/25 both eyes with correction.	eyes with correction.	Ophthalmia	C0014236	dym	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	PHYSICAL EXAMINATION	On examination visual acuity is 20/25 both eyes with correction.	20/25	20/25	C1690939	findg	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	HISTORY OF PRESENT ILLNESS	For medications, past medical, family, and social history, review of syste...	social history.	Social history	C0424945	findg	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	HISTORY OF PRESENT ILLNESS	For medications, past medical, family, and social history, review of syste...	scanned documents.	scanned	C0441633	dap	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	HISTORY OF PRESENT ILLNESS	For medications, past medical, family, and social history, review of syste...	mental status.	Mental status	C0278060	findg	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	HISTORY OF PRESENT ILLNESS	He denies pain, double vision, floaters, or flashing lights.	flashing lights.	Flashing lights	C1705500	zossy	denies
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	HISTORY OF PRESENT ILLNESS	He denies pain, double vision, floaters, or flashing lights.	flashing lights.	Flashing lights	C0085635	zossy	denies
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	HISTORY OF PRESENT ILLNESS	He denies pain, double vision, floaters, or flashing lights.	floaters.	Floaters	C0016242	findg	denies
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	HISTORY OF PRESENT ILLNESS	He denies pain, double vision, floaters, or flashing lights.	double vision,	Double Vision	C0012569	dym	denies
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	HISTORY OF PRESENT ILLNESS	He denies pain, double vision, floaters, or flashing lights.	pain,	Pain	C0030193	zossy	denies
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	HISTORY OF PRESENT ILLNESS	This patient complains of decreased vision bilaterally.	Decreased vision	Decreased vision	C0558171	findg	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	HISTORY OF PRESENT ILLNESS	This is a 64-year-old male with a history of colon cancer who was referre...	a 64-year-old male with ...	Cancer of Colon	C0007102	neop	
6ad9f783ac6453d4a6a4b1070f3754	Consultation	12/10/2011	HISTORY OF PRESENT ILLNESS	This is a 64-year-old male with a history of colon cancer who was referre...	a 64-year-old male with ...	Male	C0024554	findg	

Evaluation of quality metric extraction

As explained in the method section, we calculated Precision, Recall, and F-Measure tests to evaluate the percentage agreement between our approach and the gold standard.

For each quality metric under study we sequentially calculated precision, recall, and F measure in 4 states to measure the cumulative effect of all ontological layers combined on the base NLP output. For a given quality metrics, we first performed a query, within our repository environment, looking for the root quality metric concept like Diabetes Mellitus. We captured the result of comparing the result of this query with the gold standard as the base NLP output layer and in the form of precision, recall, and F Measure values. Subsequently, we included the concept ontology in our query and once again calculated agreement measures. We executed our query two more times after adding negation context and section ontology to the previous queries and calculated agreement measures twice more. The cumulative results after addition of each layer are shown in Table 12 for each quality metric under study.

Table 12 - *Agreements statistics results after addition of each layer (cumulative) for the quality metrics extracted from narrative texts*

Quality Metric	Layer	TP	FP	FN	TN	Precision	Recall	F-Measure
Hypertension	Base NLP Output	861	482	45	698	0.64	0.95	0.77
	+ Concept Ontology	861	487	45	693	0.64	0.95	0.76
	++ Negation Context	860	327	46	853	0.72	0.95	0.82
	+++ Section Ontology	844	219	62	961	0.79	0.93	0.86
Cardiac Surgery	Base NLP Output	13	39	56	1978	0.25	0.19	0.21
	+ Concept Ontology	64	80	5	1937	0.44	0.93	0.60
	++ Negation Context	64	62	5	1955	0.51	0.93	0.66
	+++ Section Ontology	63	29	6	1988	0.68	0.91	0.78
CNS Tumor	Base NLP Output	0	0	127	1959	0.00	0.00	0.00
	+ Concept Ontology	105	220	22	1739	0.32	0.83	0.46
	++ Negation Context	105	182	22	1777	0.37	0.83	0.51
	+++ Section Ontology	104	99	23	1860	0.51	0.82	0.63
Diabetes Mellitus	Base NLP Output	203	60	24	1799	0.77	0.89	0.83
	+ Concept Ontology	204	63	23	1796	0.76	0.90	0.83
	++ Negation Context	203	59	24	1800	0.77	0.89	0.83
	+++ Section Ontology	202	53	25	1806	0.79	0.89	0.84
TIA	Base NLP Output	22	177	12	1875	0.11	0.65	0.19
	+ Concept Ontology	22	179	12	1873	0.11	0.65	0.19
	++ Negation Context	21	37	13	2015	0.36	0.62	0.46
	+++ Section Ontology	21	27	13	2025	0.44	0.62	0.51

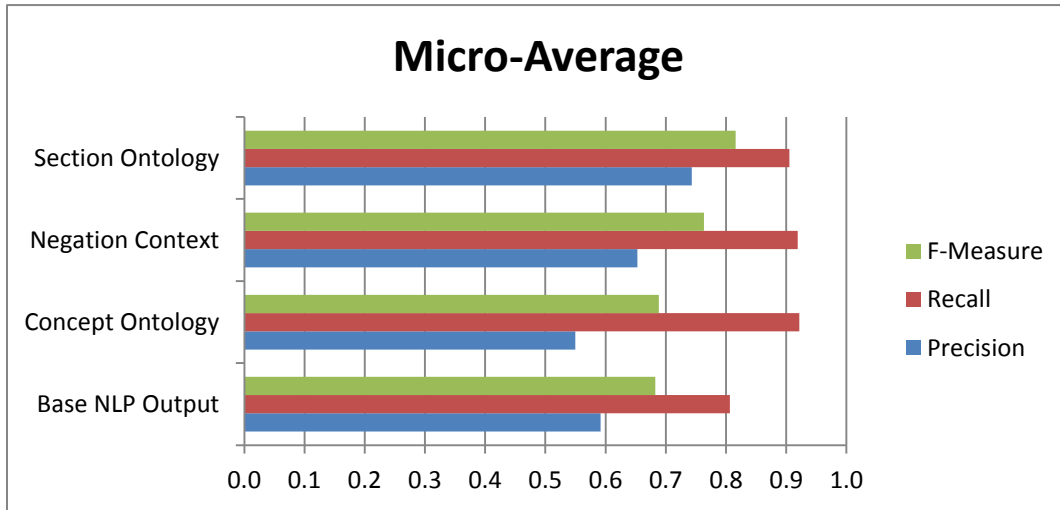
In order to calculate the combined results of all the five quality metrics we applied

Micro- averaging method (Table 13 and Figure 16)

Table 13 - *Micro-averaging the results of all 5 quality metrics combined*

	TP	FP	FN	TN	Precision	Recall	F-Measure
Base NLP Output	1099	758	264	8309	0.59	0.81	0.68
+ Concept Ontology	1256	1029	107	8038	0.55	0.92	0.69
++ Negation Context	1253	667	110	8400	0.65	0.92	0.76
+++Section Ontology	1234	427	129	8640	0.74	0.91	0.82

Figure 16 - *Micro-average combined result of agreement tests for the five quality metrics*

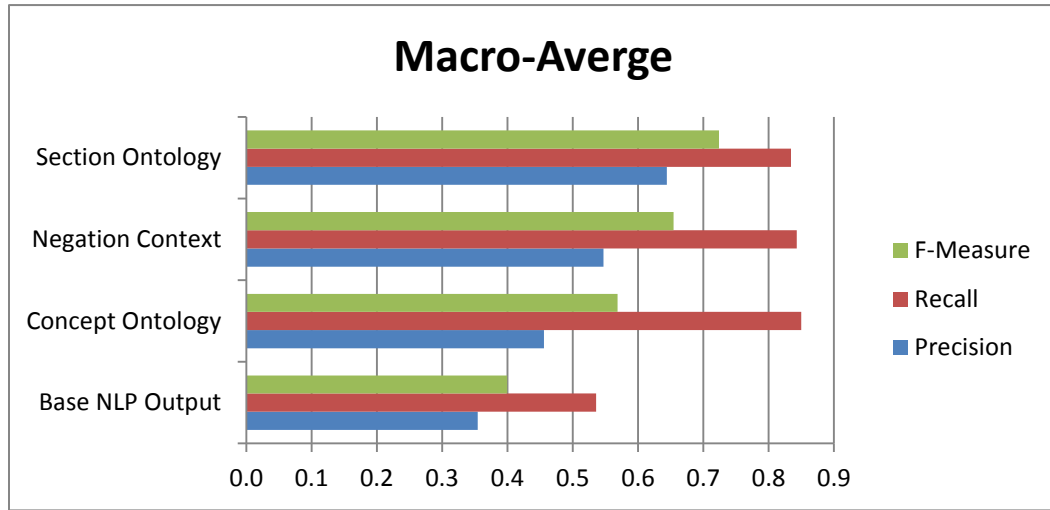


We have also looked at Macro-Averaging the results of agreement tests among the five quality metrics. Results are shown in table 14 and figure 17.

Table 14 - *Macro-averaging combined result of agreement tests for 5 quality metrics*

	Precision	Recall	F-Measure
Base NLP Output	0.35	0.54	0.40
+Concept Ontology	0.46	0.85	0.57
++Negation Context	0.55	0.84	0.65
+++Section Ontology	0.64	0.83	0.72

Figure 17 - *Macro-averaging combined agreement tests for 5 quality metrics*



In order to compare the effects of concept ontology, negation context, and section ontology on the base NLP output in isolation we computed agreement tests in a non-cumulative mode. The results of agreement tests for each layer is compared separately to the gold standard and the difference in F measure with the base NLP output is calculated. For CNS tumors there was no result for base NLP output (Table 14 & 15). Agreement test results appeared in the output only after the concept ontology is included in the query. For this reason, the calculated differences shown in Table 15 for negation context and section ontology are against concept ontology and not base NLP output. We've also combined the results of such non-cumulative comparison from all quality metrics and for each ontological layer and represented them as micro & macro averaging calculations (Table 16 & 17)

Table 15 - *Agreements statistics for each quality metric extracted from narrative texts. The difference is calculated for each layer in isolation and relative to the base NLP output*

Quality Metric	Layer	TP	FP	FN	TN	Precision	Recall	F-Measure	Diff
Hypertension	Base NLP Output	861	482	45	698	0.64	0.95	0.77	
	Concept Ontology	861	487	45	693	0.64	0.95	0.76	0.00
	Negation Context	860	323	46	857	0.72	0.95	0.82	0.06
	Section Ontology	844	216	62	964	0.79	0.93	0.86	0.09
Cardiac Surgery	Base NLP Output	13	39	56	1978	0.25	0.19	0.21	
	Concept Ontology	64	80	5	1937	0.44	0.93	0.60	0.39
	Negation Context	13	24	56	1993	0.35	0.19	0.25	0.03
	Section Ontology	13	13	56	2004	0.50	0.19	0.27	0.06
CNS Tumors	Base NLP Output	0	0	127	1959	0.00	0.00	0.00	
	Concept Ontology	105	220	22	1739	0.32	0.83	0.46	0.46
	Negation Context	105	181	22	1778	0.37	0.83	0.51	0.04
	Section Ontology	104	98	23	1861	0.51	0.82	0.63	0.17
Diabetes Mellitus	Base NLP Output	203	60	24	1799	0.77	0.89	0.83	
	Concept Ontology	204	63	23	1796	0.76	0.90	0.83	0.00
	Negation Context	202	56	25	1803	0.78	0.89	0.83	0.00
	Section Ontology	201	50	26	1809	0.80	0.89	0.84	0.01
TIA	Base NLP Output	22	177	12	1875	0.11	0.65	0.19	
	Concept Ontology	22	179	12	1873	0.11	0.65	0.19	0.00
	Negation Context	21	35	13	2017	0.38	0.62	0.47	0.28
	Section Ontology	21	25	13	2027	0.46	0.62	0.53	0.34

Figure 18 – The difference in *F* measure for each layer relative to the base NLP output

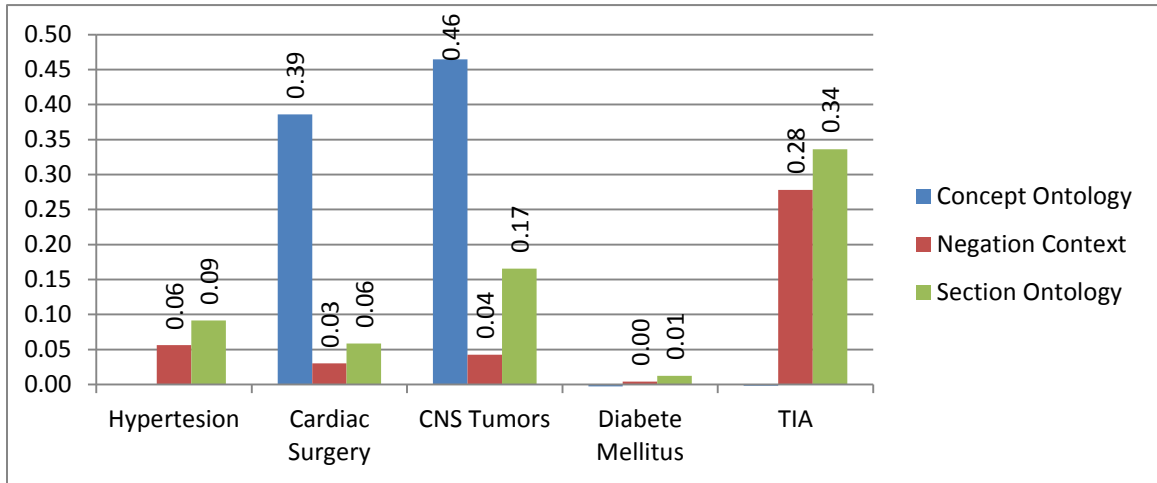


Table 16 - Micro-average result of non-cumulative agreement tests for the five quality metrics under study

	TP	FP	FN	TN	Precision	Recall	F-Measure	Diff
Base NLP Output	1099	758	264	8309	0.59	0.81	0.68	
Concept Ontology	1256	1029	107	8038	0.55	0.92	0.69	0.01
Negation Context	1201	619	162	8448	0.66	0.88	0.75	0.07
Section Ontology	1183	402	180	8665	0.75	0.87	0.80	0.12

Table 17 - Macro-average result of non-cumulative agreement tests for the five quality metrics under study

	Precision	Recall	F-Measure	Diff
Base NLP Output	0.35	0.54	0.40	
Concept Ontology	0.46	0.85	0.57	0.17
Negation Context	0.52	0.69	0.57	0.18
Section Ontology	0.61	0.69	0.63	0.23

Error Analysis

We randomly selected 10 cases of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) from each of the 5 quality metrics extraction results, except for Cardiac Surgery where only 6 false negatives were identified, and sent them to abstractors and a clinician (other than the developer) for their feedback and error analysis. All cases of TP and TN were confirmed as valid. For FP and FN cases, and upon receiving feedbacks from evaluators, we categorized responses into 7 groups (Table 18).

Table 18 – *Source of discrepancies in false positive and false negative cases*

Reason for discrepancy	%	Description
Abstractor's miss	26.0%	Quality metric valued incorrectly by abstractors
Unreachable document	16.8%	Concept found in a document outside the range of study
Concept ontology issue	15.6%	Extracted concept was not part of the concept ontology
Negation issue	12.5%	MetaMap missed the negated concept
Metric definition issue	12.5%	Metric definition was not compatible with the ontology
Section header issue	8.3%	Concept extracted from a section that was marked as irrelevant
Contextual/Uncertainty issue	8.3%	Other context dependent issues like "possible" or "questionable"
Total	100.0%	

From 25 cases of missed cases by abstractors 11 were false positives (where the correct answer by abstractors should have been “Yes”) and 14 cases were false negative (where the correct answer by abstractors should’ve been No).

In order to be compatible with abstractors, in terms of the source of the documents they reviewed during abstraction, we were tasked to look into clinical narratives (Transcribed) documented in 2011. However, during our error analysis we found that from 16 unreachable documents 4 of them were from 2010 and the rest were from sources that our

NLP engine had not access to; a database with structure documentation with no free text narratives.

There were a total of 15 of cases where the identified concept was not part of the developed ontology and therefore discarded from the results. Four cases annotated by abstractors contained Januvia, however, that drug was not included in the original reference list of the drugs used for validating diabetes mellitus (Appendix A). Two cases included deprecated SNOMED concepts and the rest (9 cases) where concepts that were located outside ontology hierarchy and within other categories. For example, *cavernous hemangioma*, which is considered as a brain tumor by abstractors, is classified under *vascular* system in SNOMED and not under *nervous system* where the ontology was built from.

During our research period we were communicating with NIH MetaMap developer team continuously and providing them with our feedback in terms of MetaMap performance, bugs, and possible enhancements. The issues we discovered in MetaMap NegEx, an algorithm for negation identification within MetaMap, were reported to NIH and validated. As a result, a new MetaMap version is expected to be released in September, 2013 that will include enhancements in negation identification that we've discovered during our analysis period.

Another discrepancy that we found in the results was ontology definition issue. In our concept ontology, per SNOMED hierarchy, *Balloon Angioplasty of Coronary Arteries* is a subtype of Heart Operation. However, abstractors consider this operation as a kind of *Percutaneous Transluminal Coronary Angioplasty* (PTCA) and counted it toward a

different metric (PTCA) in NSQIP forms. Also, abstractors did not consider a patient as hypertensive if hypertension is qualified by *mild* (mild hypertension). We categorized these instances as issues in metric definition (12 cases in the sample set).

There were 8 other cases in our error analysis samples where an extracted concept was captured under a section header (or a synonym of a section header) that was originally marked *irrelevant* in our section header ontology. Analysis of these sections and their contents showed a correctly identified concept was excluded from the results because the “irrelevant” parent section was labeled as “*Referring Physician*”, “*Specimens sent*”, or “*Attending Physician*”. In these instances, extracted (and section-less) concepts were documented in paragraphs trapped between two section headers and the upper (and irrelevant) one was flagged as the context.

The last category of discrepancy includes 8 cases where a context dependent or uncertainty concept was mistakenly valued as positive in our system. Examples include: “*elevated blood pressure only in clinic*”, “*she states that normally when she goes to a clinic, she does have elevated blood pressure readings there*”, “*possible hypertension*”, and “*questionable hypertension*”.

Limitations

We have selected a limited number of quality metrics (with simple definitions) and only from NSQIP quality metric programs for the purpose of our study. More complex quality metrics and metric from other quality collection programs may require additional pre or post processing rules and pose further challenges in information extraction algorithms.

We have also evaluated these metrics within MD Anderson transcription databases.

Analysis of clinical narratives from other healthcare organizations, with potentially different clinical narrative formats, may not necessarily results in the same results we obtained from MD Anderson environment.

Identification and extraction of subsections is a challenging task in section ontology build process. We observed that most of the subsections at MD Anderson Cancer Center were defined as part of Physical Examination section header. Therefore, we didn't outline any rule or requirements for identification and extraction of subsections in our pipeline since the source of all selected quality metrics under our study were defined by subject matter experts outside the Physical Examination section.

MetaMap

There are some challenges and limitations that should be taken into consideration when using NLP solutions for annotation of clinical text. These challenges are presented elsewhere in this paper. This section deals with the limitations of MetaMap as a practical example of the use of an integrated annotation solution in healthcare informatics. As it was mentioned earlier, acronyms and abbreviations (AA) are used frequently in the biomedical domain, specifically in clinical documentation. Once the acronym is defined, the subsequent references to the acronym will not repeat the definition. In medical context, as the sender and receiver are expected to share the common knowledge of the definition of AA terms, the acronyms are usually present without any definition. It is also noteworthy that in many situations, a specific AA can be found to have two different meanings in two different domains. MetaMap already has a set of rules to deal with this

situation, but it needs further refinement and enhancing (Aronson & Lang, 2010). Non-standard input is the other major issue when processing clinical notes. There is no consistency on the format and structure of a clinical note. Even though there are general guidelines for clinical notes (For example, the Subjective Objective Assessment Plan standard format), they are seldom followed in a consistent manner even by the same individual. To demonstrate the magnitude of this issue, Aronson and Lang found 50,000 instances of non-standard texts which resulted in false negative just for the end of sentence detection algorithm in the PubMed database. In view of the fact that the PubMed database contains text that has been carefully reviewed for publication in medical journals, this number is surprisingly high (Aronson & Lang, 2010). It is therefore prudent to assume non-standard input when working with raw clinical data obtained from patient encounters at the point of care setting.

Chapter 5: Conclusions, Discussions, and Future Directions

Recent trends in health care information systems show an increase in requirements for reporting of quality metrics by health care organizations, specifically for the government mandated programs with huge financial incentives. Healthcare providers consider EMR the best source for extracting patient information because it accurately reflects the process of patient care. Nevertheless, such valuable source of data is narrative in format, hence, inaccessible for research, unstructured for automated applications, and highly costly and time consuming for extraction by clinical abstractors.

For example, 115,000 patients were seen at MD Anderson Cancer Center in 2012, Houston, Texas ("Facts and History - Quick Facts 2013 | MD Anderson Cancer Center," 2013). Assuming for each patient visit at MD Anderson 10 narrative texts were generated 1,150,000 narrative texts were added, at minimum, to the MD Anderson Cancer EMR system in 2012 alone. Information extraction systems such as NLP solutions can be used for extraction of structured medical data from such narrative text. Although processing of clinical text is complex, effective systems have become a reality.

The availability and extension of rich knowledge bases and meta-thesaurus such as UMLS facilitates the improvement of information extraction systems and increase the use and demand for both current and historic data about the patient health profiles will drive the research for better and more efficient solutions.

The current “standard” NLP systems perform at the lexical or statistical layers of the clinical narratives; however, the embedded semantic layers should also be addressed properly in order to enhance the efficiency of such systems.

Our study introduced a framework that may contribute to advances in “complementary” components for the existing information extraction systems. The application of an ontology-based approach for natural language processing in our study has provided mechanisms for increasing the performance of such tools. The pivot point for extracting more meaningful quality metrics from clinical narratives is the abstraction of contextual semantics hidden in the notes. We have defined some of these semantics and quantified them in multiple layers in order to demonstrate the importance and applicability of an ontology-based approach in quality metric extraction. The application of such ontology layers introduces powerful new ways of querying context dependent entities from clinical texts.

It is apparent that the effect of ontology layers on information retrieval metrics (precision, recall, F measure) is largely dependent on the type of the extracted quality metric entity. Our study shows ontology layers added to the base NLP output, in general, had an increased effect of up to 63% to the performance. This effect was highest for CNS Tumors, Cardiac Surgery, and TIA concepts (63%, 57 %, 32% cumulative increase in F Measure respectively) and lowest for Hypertension and Diabetes (9% & 1 % respectively) which could be due to the format of representation of these concepts, during narration, within the clinical texts. Also, we were able to show and compare the effects of each ontology and context layer in isolation to the base NLP output. It seems section

ontology has greater effect on the overall F measure increase compared to Negation context and concept ontology on all quality metrics except for CNS Tumors and Cardiac Surgery. On a micro-average level, for all the 5 concepts combined, section header shows 11% and 5% higher values when compared to the concept ontology and negation context respectively.

Our ontology based framework achieved an overall 0.82 F Measure (micro-average) which could be suffice to be concerned, at minimum, as a decision support tool for abstractors considering the 26% missed cases we showed in the error analysis. Based on the importance of tolerable false positives or false negatives rates, for a given information extraction task, this framework can be considered as an introductory or complementary abstraction method and significantly reduces abstractor's time for extracting quality metrics hidden in the clinical narratives.

A very beneficial side effect of using such framework is the extraction of *coded and standardized* quality metric concepts which makes it a prefect process for populating structured data in clinical warehouses. Such structured, and unambiguous, concepts can also be used for explicit benchmarking, cohort studies and other data analytics where coded data is vital.

Conclusions

Reliable information about the process of care and patient outcomes is critical in correct management of healthcare services, selecting research, assurance of quality, and allocation of resources.

We have developed a framework that is necessary to identify relative semantics within

clinical text and extract a more meaningful and unambiguous quality metrics.

Furthermore, by providing bindings to standard terminologies like SNOMED CT the current approach would help quality metric extraction process becomes more objective in nature and expose data for benchmarking in a more standard way.

We believe that semantic modeling, and in particular an ontological approach, toward knowledge modeling and information extraction of quality metrics from clinical narratives can provide a unique way of improving the clarity of meaning, by providing necessary layers of disambiguation, for both human and computational systems. The use of ontology in information extraction system increases the expressivity control of extraction and helps to disambiguate retrieved concepts. This study illustrates the importance of the “complementary” role of ontologies in the existing natural language processing tools and how they can increase the general performance of quality metrics extraction task.

Rigorous evaluations are still necessary to ensure the quality of these “complementary” NLP systems. Moreover, research is needed for creating and updating evaluation guideline and criteria for assessment of the performance and efficacy of ontology-based information extraction in healthcare and to provide a consistent baseline for the purpose of comparing alternative approaches.

Future Directions

Currently, we are working on a machine learning component for this framework in order to automate section header identification and classification within the section header

ontology development process. This component will extract section headers from clinical narratives and classify sections within the ontology as soon as the clinical notes are added to the EMR system. A curator will facilitate classification for un-recognized section headers in this process. We have also including sub-section identification components which will be of benefit for concept extraction related to Physical Examination sections. Due to the high volume of clinical narratives being added each month to the MDA EMR system (~100,000 documents) and numerous sub-specialized departments at MD Anderson we anticipate creation of domain specific ontologies would be of high value for information extraction and data transactions.

We are investigating creating a certainty score to the extracted concepts based on the MetaMap score, weighted note type score, and weighted section headers score through concept frequency counts. Such certainty score could be used for filtering the results of the queries where higher degrees of performance or accuracy are needed.

An area of high interest to providers has always been problem lists within patient notes. Our approach could be used for extraction of patient problem list and the results be compared to the current IBM-cTAKES dictionary-based method.

Another area of interest in clinical studies is patient identification through cohort explorers. There may be high value in our approach for such tasks and we have proposed our framework to be included in the existing pipelines for extraction of comorbidities that will be used for such clinical trials. Last but not least, we will be looking into pattern recognition, semantic relation labeling, and knowledge discovery (through inclusion of

SemRep ontology) and including in the current proposed framework and semantic queries.

References

- About ACS NSQIP | ACS NSQIPACS NSQIP. (2013).
 AllegroGraph RDFStore Web 3.0's Database. (2013). 2013.
 Allemang, D., & Hendler, J. A. (2008). *Semantic web for the working ontologist: modeling in RDF, RDFS and OWL*: Morgan Kaufmann.
 Aronson, A. R., & Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3), 229-236. doi: 10.1136/jamia.2009.002733
 Baud, R., Lovis, C., Rassinoux, A.-M., Michel, P.-A., & Scherrer, J.-R. (1997). *Extracting linguistic knowledge from an international classification*. Paper presented at the Proceedings of MIE.
 Baud, R. H., Rassinoux, A. M., & Scherrer, J. R. (1992). Natural language processing and semantical representation of medical texts. *Methods Inf Med*, 31(2), 117-125.
 Bianchi, S., Burla, A., Conti, C., Farkash, A., Kent, C., Maman, Y., & Shabo, A. (2009). Biomedical data integration - capturing similarities while preserving disparities. *Conf Proc IEEE Eng Med Biol Soc*, 2009, 4654-4657. doi: 10.1109/IEMBS.2009.5332650
 Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1), D267.
 Bodenreider, O. (2006). Lexical, terminological and ontological resources for biological text mining. *Text mining for biology and biomedicine*.
 Bodenreider, O., & Burgun, A. (2009). Desiderata for an ontology of diseases for the annotation of biological datasets. *International Conference on Biomedical Ontology*.
 Brailer, D. J. (2005). Interoperability: the key to the future health care system. *Health Aff (Millwood)*, Suppl Web Exclusives, W5-19-W15-21. doi: 10.1377/hlthaff.w5.19
 Brank, J., Grobelnik, M., & Mladenic, D. (2005a). D1. 6.1 Ontology Evaluation. *Deliverable D*, 6, 2003-506826.
 Brank, J., Grobelnik, M., & Mladenic, D. (2005b). *A survey of ontology evaluation techniques*.
 Brinkley, J. F., Suci, D., Detwiler, L. T., Gennari, J. H., & Rosse, C. (2006). A framework for using reference ontologies as a foundation for the semantic web. *AMIA Annu Symp Proc*, 96-100. doi: 85892
 Brook, R. H., Chassin, M. R., Fink, A., Solomon, D., Kosecoff, J., & Park, R. E. (1991). A Method for the Detailed Assessment of the Appropriateness of Medical Technologies.

- Burgun, A., Golbreich, C., & Jacquelinet, C. (2004). Evolving from standard vocabularies to formal ontology for an information system dedicated to organ transplantation. *Stud Health Technol Inform*, 102, 132-144.
- Carlson, D., Farkash, A., & Timm, J. T. (2010). A model-driven approach for biomedical data integration. *Stud Health Technol Inform*, 160(Pt 2), 1164-1168.
- Ceusters, W., Lovis, C., Rector, A., & Baud, R. (1996). *Natural language processing tools for the computerised patient record: present and future*. Paper presented at the Toward an Electronic Health Record Europe.
- Ceusters, W., Smith, B., & Flanagan, J. (2003). *Ontology and medical terminology: why description logics are not enough*.
- Ceusters, W., Smith, B., & Goldberg, L. (2005). A terminological and ontological analysis of the NCI thesaurus. *Methods of Information in Medicine*, 44(4), 498-507.
- Ceusters, W., Smith, B., Kumar, A., & Dhaen, C. (2004). Ontology-based error detection in SNOMED-CT. *Stud Health Technol Inform*, 107(Pt 1), 482-486. doi: D040004186
- Chassin, M. R. (1997). Assessing strategies for quality improvement. *Health Aff (Millwood)*, 16(3), 151-161.
- Chassin, M. R., & Galvin, R. W. (1998). The urgent need to improve health care quality. Institute of Medicine National Roundtable on Health Care Quality. *JAMA*, 280(11), 1000-1005. doi: jst80006
- Chassin, M. R., Kosecoff, J., Park, R. E., Winslow, C. M., Kahn, K. L., Merrick, N. J., . . . Brook, R. H. (1987). Does inappropriate use explain geographic variations in the use of health care services? A study of three procedures. *JAMA*, 258(18), 2533-2537.
- Chong, Q., Marwadi, A., Supekar, K., & Lee, Y. (2003). Ontology based metadata management in medical domains. *Journal of Research and Practice in Information Technology*, 35(2), 139-154.
- Christensen, L. M., Harkema, H., Haug, P. J., Irwin, J. Y., & Chapman, W. W. (2009). *ONYX: a system for the semantic analysis of clinical text*. Paper presented at the Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing.
- Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med*, 37(4-5), 394-403. doi: 98040394
- Coiera, E. (2003). *Guide to health informatics*: Hodder Arnold London:.
- Committee on Identifying and Preventing Medication Errors. Institute of Medicine. (2006). *Preventing medication errors*. Washington, DC: Natl Academy Press.

- Committee on Quality of Health Care in America. Institute of Medicine. (2000). *To err is human: building a safer health system*. Washington, DC: National Academy Press.
- Committee on Quality of Health Care in America. Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*: National Academy Press Washington, DC.
- Committee on Redesigning Health Insurance Performance Measures Payment and Performance Improvement Programs. Institute of Medicine. (2007). *Rewarding provider performance : Aligning incentives in Medicare*. Washington, DC: National Academies Press.
- Committee on Redesigning Health Insurance Performance Measures Payment and Performance Improvement Programs. Institute of Medicine. (2006). *Performance measurement : accelerating improvement*. Washington, DC: National Academies Press.
- Committee to Design a Strategy for Quality Review and Assurance in Medicare. Institute of Medicine. (1990). *Medicare : a strategy for quality assurance* (Vol. 1). Washington, D.C.: National Academy Press.
- Dale, R., Moisl, H. L., & Somers, H. L. (2000). *Handbook of natural language processing*: CRC Press.
- Deming, W. E. (2000). *Out of the crisis* (1st MIT Press ed.). Cambridge, Mass.: MIT Press.
- Denny, J. C., Miller, R. A., Johnson, K. B., & Spickard III, A. (2008). *Development and evaluation of a clinical note section header terminology*. Paper presented at the AMIA Annual Symposium proceedings.
- Denny, J. C., Spickard III, A., Johnson, K. B., Peterson, N. B., Peterson, J. F., & Miller, R. A. (2009). Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association*, 16(6), 806-815.
- Donabedian, A. (1966). Evaluating the quality of medical care. *Milbank Mem Fund Q*, 44(3), Suppl:166-206.
- Donabedian, A. (1980). *The definition of quality and approaches to its assessment*. Ann Arbor, Mich.: Health Administration Press.
- Facts and History - Quick Facts 2013 | MD Anderson Cancer Center. (2013).
- Ferranti, J. M., Musser, R. C., Kawamoto, K., & Hammond, W. E. (2006). The clinical document architecture and the continuity of care record: a critical analysis. *J Am Med Inform Assoc*, 13(3), 245-252. doi: 10.1197/jamia.M1963
- Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 651-659.

- Friedman, C., Cimino, J. J., & Johnson, S. B. (1993). A conceptual model for clinical radiology reports. *Proc Annu Symp Comput Appl Med Care*, 829-833.
- Friedman, C., & Hripcsak, G. (1998). Evaluating natural language processors in the clinical domain. *Methods Inf Med*, 37(4-5), 334-344.
- Friedman, C., & Hripcsak, G. (1999). Natural language processing and its future in medicine. *Academic Medicine*, 74(8), 890-895.
- Geurts, B., & Beaver, D. I. (2011). Discourse representation theory.
- Hahn, U., & Schulz, S. *Towards a broad-coverage biomedical ontology based on description logics*. 2003.
- Harris, M. A. (2008). Developing an ontology. *Methods Mol Biol*, 452, 111-124. doi: 10.1007/978-1-60327-159-2_5
- Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing*, 32(4), 1008-1015.
- Hewitt, M. E., & Simone, J. V. (1999). *Ensuring quality cancer care*: National Academies Press.
- Hripcsak, G., Friedman, C., Alderson, P. O., DuMouchel, W., Johnson, S. B., & Clayton, P. D. (1995). Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med*, 122(9), 681-688.
- Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3), 296-298.
- Hripcsak, G., & Wilcox, A. (2002). Reference Standards, Judges, and Comparison Subjects Roles for Experts in Evaluating System Performance. *Journal of the American Medical Informatics Association*, 9(1), 1-15.
- Humphreys, B. L., & Lindberg, D. (1992). The unified medical language system project: a distributed experiment in improving access to biomedical information. *Methods Inf Med*, 7(2), 1496-1500.
- Hung, P. W., & Stetson, P. D. (2007). Development of a quality measurement ontology in OWL. *AMIA Annu Symp Proc*, 984.
- Ingenerf, J., Reiner, J., & Seik, B. (2001). Standardized terminological services enabling semantic interoperability between distributed and heterogeneous systems. *Int J Med Inform*, 64(2-3), 223-240. doi: S1386505601002118
- Johnson, P. D., Tu, S. W., Musen, M., & Purves, I. (2001). *A virtual medical record for guideline-based decision support*.
- Juran, J. M. (2004). *Architect of quality : the autobiography of Dr. Joseph M. Juran*. New York: McGraw-Hill.
- Kamal, J., Borlawsky, T., & Payne, P. R. (2007). Development of an ontology-anchored data warehouse meta-model. *AMIA Annu Symp Proc*, 1001.

- Kamp, H., Van Genabith, J., & Reyle, U. (2011). Discourse representation theory *Handbook of philosophical logic* (pp. 125-394): Springer.
- Kaplan, D. (1979). Logic of Demonstratives. *Journal of Philosophical Logic*, 8(1), 81-98.
- Kavanagh, P. L., Adams, W. G., & Wang, C. J. (2009). Quality indicators and quality assessment in child health. *Arch Dis Child*, 94(6), 458-463. doi: 10.1136/adc.2008.137893
- Kim, J. J., & Rebholz-Schuhmann, D. (2011). Improving the extraction of complex regulatory events from scientific text by using ontology-based inference. *J Biomed Semantics*, 2 Suppl 5, S3. doi: 10.1186/2041-1480-2-S5-S3
- Lambrix, P., Tan, H., Jakoniene, V., & Strömbäck, L. (2007). *Biological ontologies. Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*: Springer.
- Landgrebe, J., & Smith, B. (2011). International Conference on Biomedical Informatics. Retrieved 08/01/2011, from http://icbo.buffalo.edu/2011/slides/Friday%20Session2/Landgrebe_hl7SAIF.pdf
- Leavitt, M. (2008). Quality Measurement and Public Reporting in the Current Health Care Environment. *National Committee on Vital and Health Statistics*. Retrieved 02/18/2011, 2011, from <http://www.ncvhs.hhs.gov/080128lt.pdf>
- Lee, W. N., Tu, S. W., & Das, A. K. (2009). Extracting cancer quality indicators from electronic medical records: evaluation of an ontology-based virtual medical record approach. *AMIA Annu Symp Proc*, 2009, 349-353.
- Lewis, D. (1972). General Semantics. *Philosophy of Science*, 39(1), 111.
- Li, Y., Lipsky Gorman, S., & Elhadad, N. (2010). *Section classification in clinical notes using supervised hidden markov model*. Paper presented at the Proceedings of the 1st ACM International Health Informatics Symposium.
- Madani, S., Sittig, D., & Riben, M. (2010). Building a framework for Semantic enrichment of a Clinical note ontology. *AMIA Proceedings*.
- Magoutas, B., Halaris, C., & Mentzas, G. (2007). An ontology for the multi-perspective evaluation of quality in e-government services. *Electronic Government, Proceedings*, 4656, 318-329.
- McCray, A. T., & Nelson, S. J. (1995). The representation of meaning in the UMLS. *Methods Inf Med*, 34(1-2), 193-201.
- McDonald, C. J. (1997). The barriers to electronic medical record systems and how to overcome them. *J Am Med Inform Assoc*, 4(3), 213-221.
- McGlynn, E. A., Asch, S. M., Adams, J., Keeseey, J., Hicks, J., DeCristofaro, A., & Kerr, E. A. (2003). The quality of health care delivered to adults in the United States. *N Engl J Med*, 348(26), 2635-2645. doi: 10.1056/NEJMsa022615

- McGuinness, D. L., & Van Harmelen, F. (2004a). OWL web ontology language overview. *W3C recommendation*. Retrieved 01/16/2011, from <http://www.w3.org/TR/owl-features/>
- McGuinness, D. L., & Van Harmelen, F. (2004b). OWL web ontology language overview. *W3C recommendation*, 10, 2004-2003.
- Medical Records, Coding & Health Information Management: AHIMA Facts. (2013).
- Melton, G. B., & Hripcsak, G. (2005). Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association*, 12(4), 448-457.
- Meystre, S. M., Thibault, J., Shen, S., Hurdle, J. F., & South, B. R. (2010). Texttractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. *Journal of the American Medical Informatics Association*, 17(5), 559-562.
- Milewski, R. J., Govindaraju, V., & Bhardwaj, A. (2009). Automatic recognition of handwritten medical forms for search engines. *International Journal of Document Analysis and Recognition (IJDAR)*, 11(4), 203-218.
- Miller, R. D. (2010). *Miller's anesthesia* (7th ed. Vol. 1). Philadelphia, PA: Churchill Livingstone/Elsevier.
- Milward, D., Bjareland, M., Hayes, W., Maxwell, M., Oberg, L., Tilford, N., . . . Barnes, J. (2005). Ontology-based interactive information extraction from scientific abstracts. *Comp Funct Genomics*, 6(1-2), 67-71. doi: 10.1002/cfg.456
- Morris, C. W. (1971). Charles William Morris's Writings on the General Theory of Signs. 2013, from <http://www.angelfire.com/md2/timewarp/morris.html>
- Muir, E. The Rise and Fall of HL7. Retrieved 05/15/2011, from <https://interfaceware.fogbugz.com/default.asp?W252>
- Muir, E. (2013). HL7 Watch: The Rise and Fall of HL7. from <http://hl7-watch.blogspot.com/2011/03/rise-and-fall-of-hl7.html>
- Muller, H. M., Kenny, E. E., & Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11), e309. doi: 10.1371/journal.pbio.0020309
- National Committee for Quality Assurance. The state of health care quality 2010. Retrieved 02/02/2011, 2011, from <http://www.ncqa.org/Portals/0/State%20of%20Health%20Care/2010/SOHC%202010%20-%20Full2.pdf>
- National Quality Forum Quality Data Model. Retrieved 02/4/2011, from http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx
- National Quality Forum Quality Data Model. Retrieved 05/22/2011, from http://www.qualityforum.org/Projects/e-g/eMeasures/Electronic_Quality_Measures.aspx

- Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Retrieved 03/28/2011, 2011, from http://protege.stanford.edu/publications/ontology_development/ontology101.pdf
- Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., . . . Chute, C. G. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl 2), W170-W173.
- Obrst, L., Ashpole, B., Ceusters, W., Mani, I., & Smith, B. (2007). *The Evaluation of Ontologies: Toward Improved Semantic Interoperability*: Springer.
- Open Biological and Biomedical Ontologies: Archive of original principles. (2013). 2013, from http://www.obofoundry.org/crit_2006.shtml
- Parachoor, S. B., Rosow, E., & Enderle, J. D. (2003). Knowledge management system for benchmarking performance indicators using statistical process control (SPC) and virtual instrumentation (VI). *Biomedical Sciences Instrumentation, Vol 39*, 39, 175-178.
- Patient Handout - Diabetes Medicaiton. (2013). 2013, from http://nursing.advancweb.com/sharedresources/advancefornurses/resources/downloadableresources/n1020303_p32handout.pdf
- Pisanelli, D. M., & Gangemi, A. (2004). If ontology is the solution, what is the problem? *Ontologies in Medicine*, 102, 1-19.
- Quality Assurance Project. (2001). Cost and Quality in Healthcare. Reference Manual. *Core Training Series*. Retrieved 03/23/2011, from <http://www.hciproject.org/sites/default/files/Reference%20Manual.pdf>
- Rassinoux, A., Baud, R., & Scherrer, J. (1990). *Proximity processing of medical text*. Paper presented at the Medical Informatics Europe.
- Rassinoux, A. M., Michel, P. A., Juge, C., Baud, R., & Scherrer, J. R. (1994). Natural-Language Processing of Medical Texts within the Helios Environment. *Computer Methods and Programs in Biomedicine*, 45, S79-S96.
- Rassinoux, A. M., Wagner, J. C., Lovis, C., Baud, R. H., Rector, A., & Scherrer, J. R. (1995). Analysis of medical texts based on a sound medical model. *Proc Annu Symp Comput Appl Med Care*, 27-31.
- Raval, M. V., Bilimoria, K. Y., Stewart, A. K., Bentrem, D. J., & Ko, C. Y. (2009). Using the NCDB for cancer care improvement: an introduction to available quality assessment tools. *Journal of surgical oncology*, 99(8), 488-490.
- Ries, L. A. G. (1999). *Cancer incidence and survival among children and adolescents : United States SEER program 1975-1995*. Bethesda, MD: National Cancer Institute, SEER Program.
- Rossi Mori, A., & Consorti, F. (1998). Exploiting the terminological approach from CEN/TC251 and GALEN to support semantic interoperability of healthcare record systems. *Int J Med Inform*, 48(1-3), 111-124. doi: S1386-5056(97)00116-0

- Sager, N., Lyman, M., Nhan, N. T., & Tick, L. J. (1995). Medical language processing: applications to patient data representation and automatic encoding. *Methods Inf Med*, 34(1-2), 140-146.
- Saturno, P. J. (1999). Quality in health care: models, labels and terminology. *Int J Qual Health Care*, 11(5), 373-374.
- Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5), 507-513.
- Scherrer, J. R., Revillard, C., Borst, F., Berthoud, M., & Lovis, C. (1994). Medical Office Automation Integrated into the Distributed Architecture of a Hospital Information-System. *Methods Inf Med*, 33(2), 174-179.
- Semantic Network. (2009, 2009-09). 2013, from <http://www.ncbi.nlm.nih.gov/pubmed/>
- Shah, N., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A., & Musen, M. (2009). Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics*, 10(Suppl 9), S14.
- Shah, N. H., Jonquet, C., Chiang, A. P., Butte, A. J., Chen, R., & Musen, M. A. (2009). Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics*, 10(Suppl 2), S1.
- Shapiro, S. C. (1992). *ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE SECOND EDITION*: New Jersey: A Wiley Interscience Publication.
- Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. New York,: D. Van Nostrand Company, Inc.
- Shiloach, M., Frencher Jr, S. K., Steeger, J. E., Rowell, K. S., Bartzokis, K., Tomeh, M. G., . . . Hall, B. L. (2010). Toward robust information: data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program. *Journal of the American College of Surgeons*, 210(1), 6-16.
- SKOS Simple Knowledge Organization System - home page. (2013). 2013.
- Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., . . . Consortium, O. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11), 1251-1255. doi: 10.1038/nbt1346
- Smith, B., & Ceusters, W. (2006). HL7 RIM: an incoherent standard. *Stud Health Technol Inform*, 124, 133-138.
- Soysal, E., Cicekli, I., & Baykal, N. (2010). Design and evaluation of an ontology based information extraction system for radiological reports. *Comput Biol Med*, 40(11-12), 900-911. doi: 10.1016/j.compbiomed.2010.10.002

- Spackman, K. A., Campbell, K. E., & CÃ, R. (1997). *SNOMED RT: a reference terminology for health care*.
- SPARQL Query Language for RDF. (2013). 2013.
- Speaks, J. (2006). Truth theories, translation manuals, and theories of meaning. *Linguistics and Philosophy*, 29(4), 487-505. doi: 10.1007/s10988-006-0006-z
- Spinks, T. E., Walters, R., Feeley, T. W., Albright, H. W., Jordan, V. S., Bingham, J., & Burke, T. W. (2011). Improving cancer care through public reporting of meaningful quality measures. *Health Aff (Millwood)*, 30(4), 664-672. doi: 10.1377/hlthaff.2011.0089
- Spyns, P. (1996). Natural language processing in medicine: an overview. *Methods Inf Med*, 35(4-5), 285-301.
- Spyns, P. (2000). *Natural Language Processing in Medicine: Design, Implementation and Evaluation of an Analyser for Dutch* (Vol. 8): Leuven University Press.
- Steinberg, S. M., Popa, M. R., Michalek, J. A., Bethel, M. J., & Ellison, E. C. (2008). Comparison of risk adjustment methodologies in surgical quality improvement. *Surgery*, 144(4), 662-669.
- Stojanovic, L., Schneider, J., Maedche, A., Libischer, S., Studer, R., Lumpp, T., . . . Dinger, J. (2004). The role of ontologies in autonomic computing systems. *Ibm Systems Journal*, 43(3), 598-616.
- Tang, P. C., Ralston, M., Arrigotti, M. F., Qureshi, L., & Graham, J. (2007). Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. *J Am Med Inform Assoc*, 14(1), 10-15. doi: M2198
- Tepper, M., Capurro, D., Xia, F., Vanderwende, L., & Yetisgen-Yildiz, M. (2012). *Statistical Section Segmentation in Free-Text Clinical Records*. Paper presented at the LREC.
- Thomas, E. J., Studdert, D. M., Newhouse, J. P., Zbar, B. I., Howard, K. M., Williams, E. J., & Brennan, T. A. (1999). Costs of medical injuries in Utah and Colorado. *Inquiry*, 36(3), 255-264.
- TopQuadrant | Products | TopBraid Composer. Retrieved 01/18/2011, from http://www.topquadrant.com/products/TB_Composer.html
- Tu, S. W., Campbell, J. R., Glasgow, J., Nyman, M. A., McClure, R., McClay, J., . . . Weida, T. (2007). The SAGE Guideline Model: achievements and overview. *Journal of the American Medical Informatics Association*, 14(5), 589-598.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460.
- UMLS® Reference Manual. (2009). 2012. doi: <http://www.ncbi.nlm.nih.gov/books/NBK9676/>
- Van Asch, V. (2013). Macro-and micro-averaged evaluation measures [[BASIC DRAFT]].

- Velamuri, S. (2010). QRDA - Technology Overview and Lessons Learned. *J Healthc Inf Manag*, 24(3), 41-48.
- View Extraction - NCBO Wiki. (2013).
http://www.bioontology.org/wiki/index.php/View_Extraction
- W3C Provenance Incubator Group Wiki. Retrieved 04/15/2011, 2011, from
http://www.w3.org/2005/Incubator/prov/wiki/Main_Page
- Wager, K. A., Lee, F. W., & Glaser, J. P. (2009). *Health care information systems: a practical approach for health care management*: Jossey-bass.
- Walker, J., Pan, E., Johnston, D., Adler-Milstein, J., Bates, D. W., & Middleton, B. (2005). The value of health care information exchange and interoperability. *Health Aff (Millwood)*, Suppl Web Exclusives, W5-10-W15-18. doi: hlthaff.w5.10
- Welcome to the NCBO BioPortal | NCBO BioPortal. (2013). 2013.
- Weng, C., Gennari, J. H., & Fridsma, D. B. (2007). User-centered semantic harmonization: a case study. *J Biomed Inform*, 40(3), 353-364. doi: S1532-0464(07)00026-3
- Wimmer, M. A., Scholl, J., & Grönlund, A. (2007). *Electronic government : 6th international conference, EGOV 2007, Regensburg, Germany, September 3-7, 2007, proceedings* (1st ed.). New York: Springer.
- Wolstencroft, K., McEntire, R., Stevens, R., Taberner, L., & Brass, A. (2005). Constructing ontology-driven protein family databases. *Bioinformatics*, 21(8), 1685-1692. doi: DOI 10.1093/bioinformatics/bti158
- Zweigenbaum, P., & Courtois, P. (1998). Acquisition of lexical resources from SNOMED for medical language processing. *Stud Health Technol Inform*(1), 586-590.

Appendix A: Diabetic Medication Nursing Reference

advance
NURSES

Patient Handout

Diabetes Medication

You should know if you have type 1 or type 2 diabetes. If you have type 1, you must take insulin. If you have type 2, you may take a pill instead of or along with insulin.

Below is a list of the common drugs used to control diabetes, how they work and the common side effects.

It is important to know the name of the diabetes drug you are taking

and any of the likely side effects. You should report side effects to your doctor, nurse or diabetes educator. Look at your insulin or pill bottle to see what drug you are taking.

Compiled by Debbie G. Moore, MSN, RN, CDE, senior clinical director, clinical operations, American Healthways, Nashville, TN.

INSULIN

How it works Insulin lowers blood glucose (blood sugar). There are many different types of insulins. They differ based on onset (when the insulin begins to work), peak (when it is working the hardest), and duration of action (how long it works).

Examples

Quick-acting insulins	<ul style="list-style-type: none"> Humalog (insulin lispro) Novolog (insulin aspart)
Short-acting insulin	<ul style="list-style-type: none"> Humulin R Novolin R
Slow-acting insulins	<ul style="list-style-type: none"> Humulin N (NPH) Novolin N (NPH) Humulin L (lente) Novolin L (lente)
Long-acting insulins	<ul style="list-style-type: none"> Humulin U (ultralente) Lantus (insulin glargine)
Mixtures (2 insulins are pre-mixed)	<ul style="list-style-type: none"> Humulin 50/50 Humulin 70/30 Humalog Mix 75/25 Novolin 70/30 Novolog Mix 70/30

Side effects Low blood glucose, weight gain, allergic reaction (rare)

SULFONYLUREA

How it works These drugs cause the pancreas to make more insulin. (The drugs listed are the more common sulfonylureas prescribed.)

Examples

Generic name	Brand name
<ul style="list-style-type: none"> glimepiride glipizide glipizide glyburide glyburide glyburide 	<ul style="list-style-type: none"> Amaryl Glucotrol Glucotrol XL DiaBeta Glynase PreTab Micronase

Side effects Low blood glucose, weight gain, rash, nausea

MEGLITINIDE / D-PHENYLANINE

How it works These drugs cause the pancreas to make more insulin and act more quickly.

Examples

Generic name	Brand name
<ul style="list-style-type: none"> repaglinide nateglinide 	<ul style="list-style-type: none"> Prandin Starlix

Side effects Low blood glucose (rare)

BIGUANIDE

How it works These drugs reduce the amount of glucose that is made by the liver and helps the body better use insulin.

Examples

Generic name	Brand name
<ul style="list-style-type: none"> metformin metformin 	<ul style="list-style-type: none"> Glucophage Glucophage XR

Side effects Nausea, diarrhea, gas, loss of appetite

THIAZOLIDINEDIONE (GLITAZONE OR TZD)

How it works These drugs help the body cells better use insulin and reduce the amount of glucose that is made by the liver.

Examples

Generic name	Brand name
<ul style="list-style-type: none"> pioglitazone rosiglitazone 	<ul style="list-style-type: none"> Actos Avandia

Side effects

- Liver damage (nausea, vomiting, fatigue, dark urine, abdominal pain)
- Fluid retention/swelling
- Decrease how well some birth control pills work

ALPHA-GLUCOSIDASE INHIBITORS

How it works These drugs help keep blood sugar in target range after a meal.

Examples

Generic name	Brand name
<ul style="list-style-type: none"> acarbose miglitol 	<ul style="list-style-type: none"> Precose Glyset

Side effects Gas, bloating, diarrhea, stomach pain

COMBINATION DRUGS

How it works Sometimes several drugs are combined and sold as one pill. The action is based on the two drugs that have been combined.

Examples

Generic name	Brand name
<ul style="list-style-type: none"> glyburide & metformin glipizide & metformin rosiglitazone & metformin 	<ul style="list-style-type: none"> Glucovance Metaglip Avandamet

Side effects Because you are taking a drug that combines two medications it is possible you will have side effects from both types of drugs. These can include nausea, low blood sugar, weight gain, rash, diarrhea, excess gas, loss of appetite, liver damage, fluid retention/swelling.

The purpose of this patient education handout is to further explain or remind you about a medical condition. This handout is a general guide only. If you have specific questions, be sure to discuss them with your health care provider. This handout may be reproduced for distribution to patients.

© 2010 American Diabetes Association. All rights reserved. This handout is for your patient's use only.

93