Quality of Care Metric Reporting from Clinical Narratives: Assessing Ontology Components

Sina Madani
Department of Clinical Analytics & Informatics
University of Texas, MD Anderson Cancer Center
Houston, TX, USA
ahmadani@mdanderson.org

Reza Alemy
School of Health Information Science
University of Victoria
Victoria, BC, Canada
alemy@uvic.ca

Dean F. Sittig, Hua Xu
School of Biomedical Informatics
University of Texas Health Science Center at Houston
Houston, TX, USA

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 extraction and reporting quality of care metrics that are interpretable and comparable across healthcare institutions.*

Keywords**—***ontology; information extraction; quality of care metric; clinical narratives*

I. INTRODUCTION

The Institute of Medicine reports a growing demand in recent years for quality improvement within the healthcare industry[1]. In response, numerous organizations have been involved in the development and reporting of quality of care measurement metrics. However, the quality metrics development process is subjective in nature [2] and competing interests exist among stakeholders. As a result, conflicting data definitions from different sources shift the burden of accurate and reliable quality of care metrics extraction and reporting to the healthcare providers [3, 4]. Furthermore, manual abstraction of quality of care metrics [4], diverse implementation of Electronic Health Record (EHR) Systems [4, 5], and the lack of standards for integration across disparate clinical and research data sources [6] deepens the complexity of consistent, valid, explicit, and comparable quality of care extraction and reporting tasks within healthcare provider organizations.

The current "standard" information extraction 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. It has been shown in non-healthcare related fields that semantic modeling and ontological approaches can be used effectively for interoperability operations among diverse environments [7].

Development and application of ontologies in the domain of quality measurements have recently become the focus of some researchers. Lee et al.[8] evaluated a Virtual Medical Record (VMR) [9] method within the Standard-Based Sharable Active Guideline Environment (SAGE)[10] 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. In another short study by Hung [11] ontological modeling was evaluated for National Quality Forum's endorsed cardiovascular 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. [12] 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 an ontology-driven approach and manually created rules. Performance-wise, they only evaluated class relationships extracted from the narrative texts.

The real meaning of a concept is relative to the context in which the concept is expressed and, therefore, can be represented in different ways in a given ontology. Identification of such contexts and their representational variations in expression and providing equivalencies among such representations are crucial tasks in any knowledge modeling and information extraction activity, especially in clinical expressions where contexts are defined mostly by section headers (like Family Medical History or Assessment).

While transcription departments in relatively large hospitals tend to follow standards for documenting section headers, healthcare providers are often allowed to create their own versions of section headers in clinical notes. Denny et al. [13] trained a classifier on a dataset of 10,677 clinical notes based on boundary detection and manual annotation of section headers . He reported Precision and Recall of 95.6% and 99% respectively. In another study by Li et al. [14] a Hidden Markov Model was used for section header classification within clinical notes. They labeled sections with 15 pre-defined section header categories (like Past Medical History). The classifier achieved a per-section and per-note accuracy of 93% and 70% respectively within a dataset of 9,697 clinical notes.

The main objective of this research is to evaluate ontological components in a natural language processing (NLP) system for the purpose of unambiguous extraction of quality of care metrics. Such complementary addition to the existing information extraction system helps enterprise data integration more efficiently (time & cost) in terms of unambiguous data exchange and more objective analytics as part of the enterprise reporting system.

II. Methodologies

A. Input Data

The dataset that we received from MD Anderson (MDA) Quality Engineering Department included the National Surgical Quality Improvement Program (NSQIP) data elements abstracted from 2,085 patients who had undergone surgery in 2011. It includes a spreadsheet of quality of care metrics, such as patient's Diabetes or Hypertension, as Boolean values (Yes/No) for each patient. We considered this reported operational dataset as the gold standard for our study.

All transcribed documents of the 2,085 patients were extracted from the MDA Electronic Medical Record (EMR) repository (46,835 notes). Python scripting was used to eliminate unwanted characters and extract section headers. A typical clinical note is composed of regions of texts. Each region consists of a section header (like Chief Complaint, History of Present Illness, Physical Exam, etc.) and the relevant content in free text format.

B. Metric Selection

Abstractors at MDA abstract and report quality of care metrics in the preoperative risk assessment section of the form and send them to NSQIP. We have selected the top 5 of these variables in terms of frequency of positive cases (Boolean value="Yes") among our gold standard and for the purpose of our research. These metrics include Diabetes Mellitus, Hypertension, Transient Ischemic Attack (TIA), Cardiac Surgery, and Nervous System Tumor.

Quality of care metrics are generally documented by physicians in clinical notes. Abstractors have to read such notes and manually extract and report them to NSQIP. It should be mentioned that abstractors are nursing staff who have extensive training in NSQIP abstraction protocols & guidelines. They are also actively participating in NSQIP certification, auditing, and training programs. Shiloach et al. [15] looked into inter-rater reliability metrics and found a 1.56% disagreement rate among abstractors of the participating hospitals in NSQIP program. NSQIP data also shows that reliability has been improved with continuous training and auditing since the start of the program in 2005.

C. Natural Language Processing Engine

We implemented the National Institute of Health natural language processing engine (MetaMap v2012) [16] that is available for free for research community. A Python script pulled clinical notes from EMR repository and submit the text content of each section header, for any given clinical note, to the MetaMap for NLP analysis. In order to reduce the noise in the output we limited MetaMap processing options to RxNorm & SNOMED terminologies, minimum evaluation score of 580, and certain Unified Medical Language System semantic group (Disorders) and semantic type (Pharmacologic Substance) [17]. One XML file was generated for each note (46,835 totals) and contained patient encrypted metadata and the NLP results of the section header contents of the note.

D. Data Format and Repository Type

In order to decrease the size of the XML data obtained from the previous step we pruned unwanted XML elements from MetaMap's output. Subsequently, we converted the XML files into a RDF format and loaded them into a local instance of AllegroGraph® repository. We also used SPARQL Protocol and RDF Query Language [18] to perform federated queries across different ontologies and the RDF repository (Figure I)

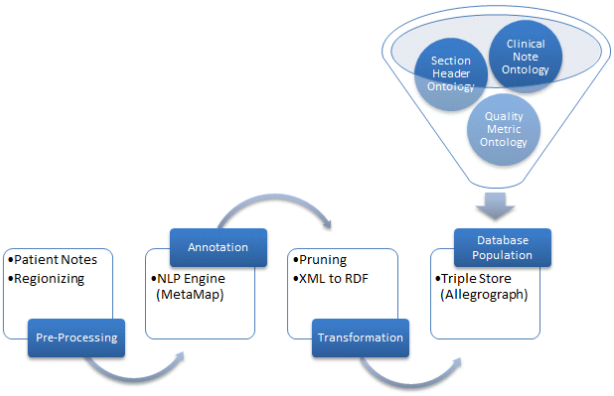

FIGURE I. NLP PIPELINE & ONTOLOGY COMPONENTS

III. Results

A. Section Header Ontology

In order to evaluate our section header extraction algorithm we randomly selected 500 test notes (100 notes from each identified quality of care metric category) and evaluated for Precision and Recall. Notes were examined by subject matter experts, annotated for section headers, and compared to the automated section header extraction algorithm. Precision, Recall, and F-measure were calculated as 99%, 97%, and 98% respectively.

In order to build our section header ontology from all extracted section headers we used SKOS *narrower* and *broader* properties for classifying section headers into hierarchies and *closeMatch*, and *exactMatch* properties [19] for assigning synonyms. After getting feedback from subject matter experts and for SPARQL query purposes each section header was categorized as relevant (like Assessment, Medical History, or Impression) or irrelevant (like Family Medical History, Recommendation, or Complications).

B. Quality of Care Metric Ontology

We identified the root concept for each of the selected quality of care metrics in SNOMED terminology (Jan 2013 version) and extracted all of their children (or subtypes). The SNOMED root concepts include: Cardiac Surgery Procedure, Tumor of Nervous System, Diabetes Mellitus, Hypertension, and Transient Ischemic Attack. According to the quality of care metric definition for Diabetes Mellitus, a patient should also take a diabetes related medication in order to be reported as a diabetic patient. For this purpose, we included diabetes mellitus medications in the ontology, with mappings to RxNorm, from the same reference [20] that abstractors used to match patient medication with diabetes in their manual abstraction process. We also reviewed this 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 quality of care metric ontology.

C. Clinical Note Ontology

For this ontology we created seven main classes, together with their relationships, in Web Ontology Language: Patient, Note, Region, Utterance, Phrase, Mapping, and Negation. All 46,835 RDF instances described in the method section were imported into the clinical note ontology within AllegroGraph® repository. The number of instances and associated data type properties for each class are shown in Table 1. Including relationships in instance count, the repository contained 70,907,728 triples. We used SPARQL for filtering unwanted concepts (within quality of care metric ontology), negated concepts, and irrelevant sections (within section ontology) from our query results.

TABLE I. CLINICAL NOTE ONTOLOGY COMPONENTS

Class	**Clinical Note Ontology Components**	
	Instance count	***Data Type properties***
Patient	2,085	Patient Id
Note	46,835	Note type, date, service, id
Region	475,692	Section header text
Utterance	2,343,856	Utterance text
Phrase	11,627,224	Phrase text
Mapping	3,263,338	Semantic type, concept, code, score
Negation	535,205	Negation trigger, type, concept, code

D. Evaluation of Quality Metric Extraction

We calculated Precision (P), Recall (R), and Micro F-measure (F) to evaluate the percentage agreement between our approach and the gold standard. When there are multiple classes of contingency tables, averaging the evaluation scores provides a more general picture of all classes combined. Micro-averaging is the most common averaging method in which each extracted instance is given the same weight. For each quality of care metric under study we sequentially calculated Precision, Recall, and F-measure in 4 conditions to measure the cumulative effect of the two ontologies and the negation context on the base NLP output. For a given quality of care metric, we first performed a query and looked for the root quality metric concept like Diabetes Mellitus. We captured the result of comparing the outcome of this query with the gold standard as the base NLP output layer and in the form of Precision, Recall, and F-measure values. Then we included the quality of care metric 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 (Table II). False Positives and Negatives (FP, FN) were calculated when there was a disagreement between each query result and the gold standard.

TABLE II. MICRO-AVERAGE RESULTS AFTER ADDITION OF EACH LAYER

Layer	TP	FP	FN	TN	P	R	F
Base NLP	1099	758	264	8309	0.59	0.81	0.68
+Metric Ont	1256	1029	107	8038	0.55	0.92	0.69
++ Negation	1253	667	110	8400	0.65	0.92	0.76
+++Section Ont	1234	427	129	8640	0.74	0.91	0.82

In order to compare isolated effect of each ontology and the negation context on the base NLP output we computed agreement tests in a non-cumulative mode as well. The micro-average 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 (Table III).

TABLE III. EFFECT OF EACH ONTOLOGY LAYER ON BASE NLP OUTPUT

Layer	P	R	F	Difference with Base NLP Output
Base NLP	0.59	0.81	0.68	
Metric Ont	0.55	0.92	0.69	0.01
Negation	0.66	0.88	0.75	0.07
Section Ont	0.75	0.87	0.80	0.12

IV. DISCUSSION

Recent trends in health care information systems show an increase in requirements for reporting of quality of care 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 most accurately reflects the process of patient care. Nevertheless, such a valuable source of data is usually in narrative format, hence, inaccessible for easy structured reports, and highly costly and time consuming for manual extraction by clinical abstractors.

Our study introduced a framework that may contribute to the advances in *"*complementary*"* components for the existing information extraction systems. The application of ontology components for the NLP system in our study has provided mechanisms for increasing the performance of such tools. The pivot point for extracting more meaningful quality of care 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 a quality of care metric extraction system. The application of ontology components introduces powerful new ways of querying context dependent entities from clinical narratives.

It is apparent that the effect of ontology components on information retrieval metrics (Precision, Recall, F-measure) are largely dependent on the type of the quality of care metric. Our study shows ontology layers added to the base NLP output, in general, had an increased effect of up to 63% to the performance. The cumulative increase in F-measure was highest for Nervous System Tumors, Cardiac Surgery, and TIA (63%, 57 %, and 32% respectively) and lowest for Hypertension and Diabetes (9% & 1 % respectively) which could be due to the format of representation of these concepts within the clinical narratives. Also, we were able to show and compare the effects of each ontology and negation context in isolation to the base NLP output. It seems section header ontology has a greater effect on the overall F-measure increase compared to the negation context and quality of care metric ontology on all quality metrics except for Nervous System Tumors and Cardiac Surgery. On a micro-average level, for all the 5 concepts combined, section header ontology shows 11% and 5% higher values when compared to the quality of care metric ontology and negation context respectively.

Our ontology-based framework achieved an overall 0.82 F-measure (Micro) which may be high enough to be considered, at minimum, as a decision support tool. Based on the 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 of care metrics hidden in the clinical narratives.

V. CONCLUSION

We have developed a framework that helps identify contextual semantics within clinical text and extract more meaningful and unambiguous quality of care metrics for the patient care process. Furthermore, by providing bindings to standard terminologies (like SNOMED) the current approach would help quality of care metric extraction process become more objective in nature and deliver structured data for populating clinical warehouses, explicit benchmarking, cohort studies, and other clinical analytics where coded data is vital.

We believe that an ontological approach toward knowledge modeling and information extraction of quality of care 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 disambiguate the 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 the 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.

REFERENCES

[1] P. Maurette, and C. A. M. R. Sfa, "To err is human: building a safer health system," *Annales Francaises D Anesthesie Et De Reanimation,* vol. 21, no. 6, pp. 453-454, Jun, 2002.

[2] R. D. Miller, *Miller's anesthesia*, 7th ed., p.^pp. 81-2, Philadelphia, PA: Churchill Livingstone/Elsevier, 2010.

[3] P. C. Tang, M. Ralston, M. F. Arrigotti, L. Qureshi, and J. Graham, "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,* vol. 14, no. 1, pp. 10-5, Jan-Feb, 2007.

[4] S. Velamuri, "QRDA - Technology Overview and Lessons Learned," *J Healthc Inf Manag,* vol. 24, no. 3, pp. 41-8, Summer, 2010.

[5] C. J. McDonald, "The barriers to electronic medical record systems and how to overcome them," *J Am Med Inform Assoc,* vol. 4, no. 3, pp. 213-21, May-Jun, 1997.

[6] Q. Chong, A. Marwadi, K. Supekar, and Y. Lee, "Ontology based metadata management in medical domains," *Journal of Research and Practice in Information Technology,* vol. 35, no. 2, pp. 139-154, 2003.

[7] B. Magoutas, C. Halaris, and G. Mentzas, "An ontology for the multi-perspective evaluation of quality in e-government services," *Electronic Government, Proceedings,* vol. 4656, pp. 318-329, 2007.

[8] W. N. Lee, S. W. Tu, and A. K. Das, "Extracting cancer quality indicators from electronic medical records: evaluation of an ontology-based virtual medical record approach," *AMIA Annu Symp Proc,* vol. 2009, pp. 349-53, 2009.

[9] P. D. Johnson, S. W. Tu, M. Musen, and I. Purves, "A virtual medical record for guideline-based decision support." p. 294.

[10] S. W. Tu, J. R. Campbell, J. Glasgow, M. A. Nyman, R. McClure, J. McClay, C. Parker, K. M. Hrabak, D. Berg, and T. Weida, "The SAGE Guideline Model: achievements and overview," *Journal of the American Medical Informatics Association,* vol. 14, no. 5, pp. 589-598, 2007.

[11] P. W. Hung, and P. D. Stetson, "Development of a quality measurement ontology in OWL," *AMIA Annu Symp Proc*, pp. 984, 2007.

[12] E. Soysal, I. Cicekli, and N. Baykal, "Design and evaluation of an ontology based information extraction system for radiological reports," *Comput Biol Med,* vol. 40, no. 11-12, pp. 900-11, Nov-Dec, 2010.

[13] J. C. Denny, A. Spickard III, K. B. Johnson, N. B. Peterson, J. F. Peterson, and R. A. Miller, "Evaluation of a method to identify and categorize section headers in clinical documents," *Journal of the American Medical Informatics Association,* vol. 16, no. 6, pp. 806-815, 2009.

[14] Y. Li, S. Lipsky Gorman, and N. Elhadad, "Section classification in clinical notes using supervised hidden markov model." pp. 744-750.

[15] M. Shiloach, S. K. Frencher Jr, J. E. Steeger, K. S. Rowell, K. Bartzokis, M. G. Tomeh, K. E. Richards, C. Y. Ko, and B. L. Hall, "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*, vol. 210, no. 1, pp. 6-16, 2010.

[16] A. R. Aronson, and F. M. Lang, "An overview of MetaMap: historical perspective and recent advances," *J Am Med Inform Assoc*, vol. 17, no. 3, pp. 229-36, May-Jun, 2010.

[17] "MetaMap Semantic Groups and Types," 07/10/2014, 2014; http://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml.

[18] "SPARQL Query Language for RDF," 2014; http://www.w3.org/TR/rdf-sparql-query/.

[19] "SKOS Simple Knowledge Organization System Namespace Document - HTML Variant, 18 August 2009 Recommendation Edition," 2014; http://www.w3.org/2009/08/skos-reference/skos.html.

[20] "Patient Handout - Diabetes Medicaiton," 2013; http://nursing.advanceweb.com/sharedresources/advancefornurses/resources/downloadableresources/n1020303_p32handout.pdf.