Application of HL7/LOINC Document Ontology to a University-Affiliated Integrated Health System Research Clinical Data Repository

Yan Wang, MS[1], Serguei Pakhomov, PhD[1, 2], Justin L. Dale, BA[3],
Elizabeth S. Chen, PhD[5, 6, 7], Genevieve B. Melton, MD, MA[1, 4]
[1]Institute for Health Informatics, [2]College of Pharmacy,[3] Academic Health Center -
Information Systems, [4]Department of Surgery, University of Minnesota, Minneapolis, MN;
[5]Center for Clinical & Translational Science, [6]Department of Medicine, [7]Department of
Computer Science, University of Vermont, Burlington, VT

Abstract
Fairview Health Services is an affiliated integrated health system partnering with the University of Minnesota to establish a secure research-oriented clinical data repository that includes large numbers of clinical documents. Standardization of clinical document names and associated attributes is essential for their exchange and secondary use. The HL7/LOINC Document Ontology (DO) was developed to provide a standard representation of clinical document attributes with a multi-axis structure. In this study, we evaluated the adequacy of DO to represent documents in the clinical data repository from legacy and current EHR systems across community and academic practice sites. The results indicate that a large portion of repository data items can be mapped to the current DO ontology but that document attributes do not always link consistently with DO axes and additional values for certain axes, particularly "Setting" and "Role" are needed for better coverage. To achieve a more comprehensive representation of clinical documents, more effort on algorithms, DO value sets, and data governance over clinical document attributes is needed.

Introduction
Electronic Health Record (EHR) systems and electronic clinical documentation allow for functions such as decision support, quality assurance and clinical research. Clinical data warehouses and other infrastructures incorporating multiple data sources can enable researchers to perform cohort identification, hypothesis generation, retrospective analyses and other research functions[1]. With the rapid proliferation of locally customized documents in clinical care, there is a growing need for standard representations to enable their storage, navigation, retrieval and use.

The HL7/LOINC Document Ontology (DO) is an ontology for standardizing clinical documents with terms in a hierarchical structure to support exchange and reuse of clinical documents across institutions and different systems[2]. It is composed of five axes: *Kind of Document (KOD)* (e.g., Letter, Report, Note), *Type of Service* (*TOS*) (e.g., Consultation, Procedure, Evaluation and Management), *Setting* (e.g., Intensive Care Unit, Birthing Center*), Subject Matter Domain* (*SMD*) (e.g., Urology, Cardiovascular Disease) and *Role* (e.g., Physician, Registered Nurse, Technician). Each axis has a set of restricted values. In addition, many precoordinated value sets for DO map to LOINC codes.

In building the University of Minnesota clinical data repository as an infrastructure primarily for clinical and translational researchers in collaboration with University of Minnesota-affiliated Fairview Health Services, both legacy and current EHR system data are incorporated. The goal of this study was to represent clinical documents in a manner to facilitate their reuse by researchers for clinical studies and for filtering documents as part of the front end of our biomedical and clinical natural language processing system, BioMedICUS[3]. We hypothesized that DO would provide an adequate framework for document representation. Another hypothesis was that the intricacies of a vendor-supported EHR system with large numbers of document value sets would likely reveal certain incongruent or inconsistent aspects with the DO framework.

Background
Several researchers have explored the application of DO, mostly with document names in the inpatient setting. In a seminal study[4], Hyun *et al.* evaluated the adequacy of three versions of the DO to represent document names in the inpatient setting at Columbia University Medical Center. The authors found DO version 3 was superior to previous versions on both the level of specificity and completeness in document names as well as level of granularity of DO. Axes *SMD* and *TOS* in version 3 had value sets that needed expansion for better representation of document names.

To understand the DO in further depth and the effect of precoordination of its axes with LOINC, Dugas *et al.*[5] evaluated the coverage of LOINC codes on 86 document types from an inpatient hospital information system.

Similar to Hyun's work, the authors reported that more specific LOINC codes were necessary for better coverage. Subsequently, Chen et al.[6] explored the process of mapping document names from two large institutions to DO and identifying LOINC codes based on the mapping to further define the strengths and limitations of DO and LOINC codes for document representation. The results showed that a majority of document names can be assigned to a LOINC code with existing DO axis values but that there was often loss of information and granularity mismatch for one or more axis(es). In this study, mappings were performed for document names from the inpatient setting at one site and document attributes from a legacy system at the second site. Li et al.[7] investigated the coverage of DO on clinical document titles again in the inpatient setting from two campuses of the NewYork-Presbyterian Hospital and explored the process of using LOINC codes for exchange of documents across institutions. Similar coverage of DO and issues identified in the aforementioned studies were reported. The authors explored different combinations of DO axes for LOINC code mapping. As a small percent of local documents were reported to have an exact match to existing LOINC codes, the study showed that using LOINC codes with precoordination for document exchange might not be feasible or allow for maximal flexibility of document reuse.

In a study looking at nursing documents and their associated section headings, Hyun and Bakken[8] extracted section headings from nursing documents, identified DO components and mapped them to the LOINC semantic model. The study reported that 38% of the headings were successfully represented. The authors also found that in order to better represent nurse document components, values of the attributes of the LOINC semantic model needed to be extended. Finally, Shapiro et al.[9] analyzed the single DO axis of *SMD* on a set of document titles within the Medical Entities Dictionary at NewYork-Presbyterian Hospital. The study showed that 56% of document titles were classified as "not specified" on the DO *SMD* list. In the study, a new polyhierarchical SMD structure was created combining the values from the DO database with values from the American Board of Medical Specialties (ABMS). The resulting new structure significantly increased the coverage of *SMD* on document titles.

Methods

This study involved analysis, collection, text processing, and mapping of clinical documents stored in the clinical data repository from University of Minnesota-affiliated Fairview Health Services. The repository contains documents between 1993 and 2013 from a single electronic health record system (Epic™), as well as legacy documents from affiliate clinics for variable time periods. Fairview includes seven hospital sites and over 330 ambulatory clinic locations. Currently the repository hosts more than 66 million notes and the average daily document volume includes 133,000 document updates and 45,000 new note insertions.

Repository data items related to patient encounters, departments, providers and clinical documents were inspected to find items with values for each DO axis. Figure 1 shows data items such as "Encounter Type", "Position Type" collected from patient encounters. Each of these data items was compared to different axes of the DO and was analyzed to understand if each of the data items specified information in one or more DO axis(es). While the DO has at minimum population of the *Kind of Document* and at least one other DO axis, the distribution of axis population was analyzed, in order to understand the consequences of filtering by one or more of the DO axes based on analysis of the items and populating rates of each data item collected from the data repository. As illustrated in Figure 1, entries of department specialty, provider specialty and hospital service were mapped to the *SMD* axis.
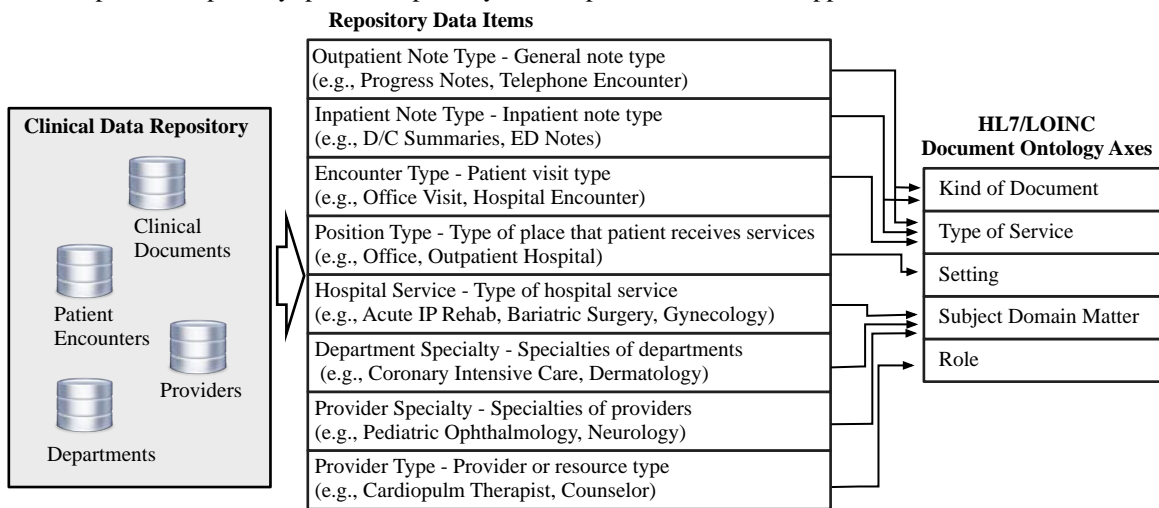
Figure 1. Mapping from Epic items to HL7/LOINC DO axis.

Table 1. Mapping examples of different data items.

Example	Document Ontology Axis(es) Mapping(s)					
Anesthesia Pre-op Evaluation (Inpatient Note Type)	KOD	7. Note	TOS	9.l.2. Preoperative Evaluation and Management	SMD	4.Anesthesiology
Consult/Results-Findings (Outpatient Note Type)	KOD	7. Note	TOS	3. Consultation		
Nursing Facility (Position Type)	Setting	6.b. Nursing Facility				
Palliative Care (Hospital Service)	SMD	30. Palliative Care				
Pediatric Neurology (Department Specialty)	SMD	19. Neurology				
Ent-Otolaryngology (Provider Specialty)	SMD	29.Otolaryngology				
Cardiopulm Therapist (Provider Type)	SMD	14.b. Cardiovascular Disease	Role	15. Therapist		
Prenatal Office Visit (Encounter Type)	TOS	7.b. Office				

Depending on the type of the visit (e.g., inpatient or outpatient) that a clinical note is associated with, different sets of data items can be used to extract information for DO axes. For example, for an inpatient encounter, the clinical document specialty (DO axis of *SMD*) usually can be extracted from the "Hospital Service", "Provider Specialty" or document titles. For an outpatient encounter, the SMD is often specified in the "Department Specialty", "Provider Specialty" or document titles. A list of data items, such as "Outpatient Note Type" and "Hospital Service", in the data repository were mapped to DO axes. Table 1 shows examples of mapping from data items to DO axes.

We adopted the same mapping and rating process described by Hyun *et al.* and Chen *et al*[4,6]. Each data item was examined and values for applicable DO axes extracted from the entry. For example, from encounter type "Oncology Visit", DO *SMD* axis is extracted as "14.f. Hematology and Oncology". If an entry included related information to a DO axis, but no appropriate value was found in the existing value list, the entry was classified as "Not Covered" for that particular DO axis. If an entry contained no information for an axis of the DO, a value of "Not Specified" was assigned to the entry. A rating (*adequate*, *too broad*, *too specific*, *not covered or not specified*) was used to indicate the coverage of the particular DO axis values to each of the above mapping results. Inter-rater reliability was calculated using mappings of two reviewers on approximately 10% random subsets of data item entries.

Results
Table 2 contains a list of data items, the number of item entries in the repository, number of entries that specify values for each DO axis, and populating rate of each data item. Entries were found to sometimes specify information for more than one DO axis. For instance, encounter type "Anesthesia Consult" indicates both the *TOS* and *SMD*. Values for each DO axis can be obtained in several data items. For instance, some item entries, such as with an encounter type of "Case Management" and Department Specialty of "Dialysis", can all indicate the *TOS* axis. Information on *SMD* was contained to some extent in all data items.

Table 2. Distinct data item mappings to DO axis values.

Data Item	Entries	KOD	TOS	Setting	SMD	Role	Populating rate
Inpatient Note Type	90	83	79	7	3	3	100% (Inpatient)
Outpatient Note Type	65	49	46	5	2	2	96.5% (Outpatient)
Position Type	50	0	0	47	0	0	74.3/98.3% (Outpatient/Inpatient)
Hospital Service	98	0	6	14	87	5	91.2% (Inpatient)
Department Specialty	95	0	2	5	82	8	26% (Outpatient)
Provider Specialty	176	0	2	5	155	28	-
Provider Type	79	0	0	0	38	71	67%/76.5% (Outpatient/Inpatient)
Encounter Type	172	57	64	40	14	6	84.9%/100% (Outpatient/Inpatient)

The distribution of populating rates of each item hosted in the data repository (with the exception of those most related to provider specialty) is also summarized in Table 2. As shown in Table 2, some data items were well populated (>80%), including encounter type, hospital service and position type; whereas others, such as the department specialty and provider type, were fairly populated. For example, 47 out of 50 entries for "Position Type"

contain information about DO axis *Setting* and 30 of them can be mapped to an existing value of DO axis *Setting*. Table 3 shows proportions of finalized inpatient and outpatient notes populated with data item entries containing DO information and data item entries can be mapped to existing DO values. 95.3% of inpatient notes are populated with "Position Type" from the 47 entries and 91.4% of inpatient notes are populated with a "Position Type" from the 30 entries. The distribution of mapping ratings for data items and inter-rater reliability of mappings are summarized in Table 4.

Table 3. Inpatient and outpatient note population rates with data item entries containing DO information and with data item entries can be mapped to existing DO values.

	Inpatient Notes n=2,134,945 (16.57%)	Outpatient Notes n=10,751,838 (83.43%)
KOD	100.0% / 100.0% (Inpatient Note Type)	96.5% / 96.5% (Outpatient Note Type)
TOS	100.0% / 34.7% (Inpatient Note Type) 100.0% / 100.0% (Encounter Type)	96.5% / 96.5% (Outpatient Note Type) 65.9% / 65.9% (Encounter Type)
Setting	95.3% / 91.4% (Position Type)	71.8% / 71.3 % (Position Type)
SMD	86.9% / 85% (Hospital Service) -/- (Provider Specialty)	16.3% / 15.8% (Department Specialty) -/- (Provider Specialty)
Role	76.2% / 76.2% (Provider Type)	59.6% / 59.6% (Provider Type)

The majority of the data item entries contain DO axis information. Overall, existing values of DO axes are either adequate or too broad for data item entries that contain DO axis information. Existing *KOD* and *SMD* values can exactly specify most of the data item entries that contain *KOD* and *SMD* information. A number of new *Setting* types such as "Community Mental Health Center" and "Independent Laboratory" as well as new *Role* types such as "Athletic Trainer" and "Diabetes Educator" were discovered from the mappings. A large number of data item entries contain more specific *TOS* and *Setting* information than existing DO values for these two axes. For example, position type "End Stage Renal Disease Treatment Facility" is mapped to a less specific *Setting* value "7. Outpatient". Outpatient note type "Consult/Results – Findings" and "Consult/Results – Impression" are mapped to the same less specific *TOS* value "3. Consultation".

Table 4. Mapping ratings by axis and inter-rater reliability.

Mapping ratings (Inpatient setting=shading; Outpatient setting=no shading)					
	KOD	TOS	Setting	SMD	Role
Adequate	87.8%	20.6%	26.0%	58.0%	29.1%
	72.3%	13.1%	26.0%	58.6%	29.1%
Too Broad	3.3%	30.1%	32.0%	11.1%	24.0%
	1.5%	27.8%	32.0%	10.3%	24.0%
Too Specific	0	0.4%	2.0%	0.8%	0
	0	0	2.0%	0.7%	0
Not Covered	1.1%	3.4%	34.0%	5.4%	36.7%
	1.5%	5.5%	34.0%	5.8%	36.7%
Not Specified	7.8%	45.4%	6.0%	24.6%	10.1%
	24.6%	53.6%	6.0%	24.6%	10.1%
Inter-Rater Reliability on SMD axis					
Proportion	88.9%	92.3%	80.0%	84.6%	85.7%
Agreement	100%	91.3%	80.0%	88.5%	85.7%

*Mapped item rates of SMD is reported for inpatient documents mapping using *hospital service* data only and for outpatient using *department specialty* only.

Discussion

In this paper, we have described an analysis applying the HL7/LOINC DO for representing documents in a large research clinical data repository for a sizable integrated health system. We studied the structure of entities related to encounters, notes, providers and departments. A set of data items were collected and then mapped to the DO axes. Database populating rate of all data item entries that mapped to a particular DO axis were calculated and analyzed for both inpatient and outpatient settings. Mapping results showed that for both inpatient and outpatient documents, the majority of the related repository item entries can be mapped to a value in the defined list of the respective DO axis. Similar to previous studies on the adequacy of DO for clinical document names, we observed similar issues such as granularity issues and loss of information.

Further analysis on the data repository shows that most inpatient notes were populated with data item entries that are mapped to existing *KOD* and *Setting* types in HL7/LOINC DO. However, only 34.7% of inpatient document was populated with data item entries that can be mapped to existing *TOS* values. Upon inspection of documents in the data repository, we found that nearly half of the inpatient documents were populated with an "Inpatient Note Type" or "Miscellaneous", which cannot be mapped to an existing *TOS* value. The populating rate of hospital service that can be mapped to DO axis *SMD* was high (86.9%). Only 59.6% of the outpatient documents are populated with provider types that can be mapped to DO axis *Role* because of the low populating rate of the provider type at the institutional level, particularly for interdisciplinary staff. Also, only 16.3% of the outpatient documents are populated with mapped department specialties, which could be used for extracting information for DO axis *SMD*.

In addition to structured data items related to *SMD*, titles of clinical documents, such as "Sleep Medicine Chart Note", "FINAL PULMONARY CONSULTATION", and "AMB Nurse Triage Note", also contain *SMD* information. For instance, the document title "Sleep Medicine Chart Note" indicates an *SMD* value "19. Neurology" and title "FINAL PULMONARY CONSULTATION" indicates an *SMD* value "14.i. Pulmonary Disease". To better utilize information encoded in document titles, algorithms and tools need to be developed in future studies.

The University of Minnesota clinical data repository stores clinical documents from the current Epic system, as well as documents from legacy EHRs (i.e., AllScripts and Eclipsys) over a sustained period of time. The process of adding documents from a number of legacy systems has resulted in inconsistencies between data item entries. For instance, analysis of documents within the repository shows that a large number of legacy documents were populated with encounter type "Admission H&P", but associated with an outpatient note type "Progress Note".

The results presented show that the HL7/LOINC DO is able to represent a majority of clinical data repository data items. Similar issues such as granularity and loss of information were found in this study as reported in previous studies. Further effort is needed to develop tools to acquire with higher fidelity *SMD* and *TOS* for documents. We plan a more detailed analysis of *Role* and *Setting*, along with providing detailed mappings to the DO of these values.

References
1. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. Journal of the American Medical Informatics Association: JAMIA. [Research Support, N.I.H., Extramural]. 2012 Jun;19(e1):e119-24.
2. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, et al. HL7 Clinical Document Architecture, Release 2. Journal of the American Medical Informatics Association: JAMIA. 2006 Jan-Feb;13(1):30-9.
3. BioMedICUS. Available from: http://code.google.com/p/biomedicus/.
4. Hyun S, Shapiro JS, Melton G, Schlegel C, Stetson PD, Johnson SB, et al. Iterative evaluation of the Health Level7--Logical Observation Identifiers Names and Codes Clinical Document Ontology for representing clinical document names: a case report. Journal of the American Medical Informatics Association: JAMIA. [Research Support, N.I.H., Extramural]. 2009 May-Jun;16(3):395-9.
5. Dugas M, Thun S, Frankewitsch T, Heitmann KU. LOINC codes for hospital information systems documents: a case study. Journal of the American Medical Informatics Association: JAMIA. 2009 May-Jun;16(3):400-3.
6. Chen ES, Melton GB, Engelstad ME, Sarkar IN. Standardizing Clinical Document Names Using the HL7/LOINC Document Ontology and LOINC Codes. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2010;2010:101-5.
7. Li L, Morrey CP, Baorto D. Cross-mapping clinical notes between hospitals: an application of the LOINC Document Ontology. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;2011:777-83.
8. Hyun S, Bakken S. Toward the creation of an ontology for nursing document sections: mapping section names to the LOINC semantic model. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. [Research Support, N.I.H., Extramural]. 2006:364-8.
9. Shapiro JS, Bakken S, Hyun S, Melton GB, Schlegel C, Johnson SB. Document ontology: supporting narrative documents in electronic health records. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. [Research Support, N.I.H., Extramural Validation Studies]. 2005:684-8.