Creating Reusable Annotated Corpora with the Clinical Document Architecture

Scott L. DuVall  
VA Salt Lake City Health Care  
System and University of Utah  
scott.duvall@utah.edu

Brett R. South  
VA Salt Lake City Health Care  
System and University of Utah  
brett.south@hsc.utah.edu

Leonard W. D’Avolio  
Massachusetts Veterans  
Epidemiology Research and Information Center  
leonard.davolio@va.gov

Keith W. Boone  
GE Healthcare, Health Level Seven,  
Integrating the Healthcare Enterprise and Health Information Technology Standards Panel  
keith.boone@ge.com

Shuying Shen  
VA Salt Lake City Health Care  
System and University of Utah  
shuying.shen@hsc.utah.edu

Adi Gundlapalli  
VA Salt Lake City Health Care  
System and University of Utah  
adi.gundlapalli@hsc.utah.edu

Jonathan R. Nebeker  
VA Salt Lake City Health Care  
System and University of Utah  
jonathan.nebeker@hsc.utah.edu

Matthew H. Samore  
VA Salt Lake City Health Care  
System and University of Utah  
matthew.samore@hsc.utah.edu

Abstract

Manual annotation of clinical documents creates reference standards to train and evaluate natural language processing (NLP) systems, but can also be used in other aspects of patient care and research. Manual annotation is costly and time-consuming and is usually done for specific use cases. We present a practical approach to the storage of annotations along with text using the Clinical Document Architecture (CDA). We describe the types of annotations commonly made on clinical text and show how they map to CDA elements. This provides standard, reusable documents that allow multiple users to modify and add annotations for new or expanded clinical use cases. As clinical texts become more widely available and NLP systems are applied to extract information from the text, annotated corpora will become an increasingly essential resource. We demonstrate how the CDA provides an appropriate mechanism for storing annotated corpora and describe how it supports interoperability and reusability.

1. Introduction

The increasing emphasis on quality measurement, pay for performance, and evidence-based medicine has led to increased demands for using patient information stored in an electronic medical record (EMR) for multiple purposes other than clinical care. Unfortunately, most EMR systems are not designed to provide information fit for each secondary data use. Because of this, a host of systems, people, and processes are employed to create, store, and manage information inside health care facilities. This paper addresses the storage and reuse of text portions of the EMR with an existing health care data standard. We show how this standard supports the integration of structure and meaning that is added to text through manual annotation and natural language processing (NLP) and describe how it provides a unifying view of information added by different people for different projects and purposes.

The paper is organized as follows: first an overview of how patient documentation is organized in an EMR is presented. The challenges and benefits of using EMR data for purposes other than patient care is discussed with an emphasis on the rich detail contained in clinical text. We describe a multi-site research consortium whose focus is developing and applying methods for extracting information from clinical texts using NLP. While the interdependencies of the different project teams in the consortium provided the motivation for paper, the ability to effectively store and reuse study data involving clinical text can benefit even the simplest studies. In addition, even though the setting for this project centers on research use of patient data, we briefly discuss how the same processes can also support automated clinical decision support, quality assurance, administrative processes, and patient care.

Next, we review current annotation methods with their limitations. We introduce the existing health care
data standard, the Clinical Document Architecture (CDA), and show the parallels in patient care and NLP research that allow for its use in storing study data. An example patient scenario is given which helps introduce the different types of annotations made on clinical text and how they are represented in the CDA. Finally, some differences in the original purpose of the CDA and its use with NLP are presented along with a road map of future work.

Our objective in presenting these ideas is to introduce the CDA as an effective way to store and reuse annotated documents, encourage the creation and adoption of tools that take advantage of the CDA for NLP studies, and generally spark discussion of the usefulness of the CDA outside of patient care.

2. Patient documentation

The patient documentation in an EMR usually consists of a series of structured elements that represent findings: observations about the patient like blood pressure, heart rate, or cholesterol levels; and diagnoses: conditions that the patient has been confirmed to have such as asthma, bronchitis, or pneumonia. Data recorded in structured format can be mapped to medical terminologies where problems, treatments, and tests are represented as alphanumerical codes. In addition to this structured information, a narrative description of the diagnosis process and the interaction with the patient is often recorded. An advantage of narrative text is that it gives clinicians autonomy in expressing their thoughts when authoring patient documentation. Using text to describe the patient’s state of health is a more natural form of recording important information than check boxes, radio buttons, and drop down lists.

2.1 Use of structured patient documentation

Much research has focused on using the structured and coded data available in patient documentation because the inherent structure of the data allows them to be used in rules and be tallied and categorized [1-2]. The elements available in structured data poorly represent some portions of the patient’s state of health. Typically only the top reimbursable diagnoses for each visit are coded, even though a clinician may be monitoring and treating dozens of problems and related conditions. Even when codes are available, these elements cannot adequately capture the specifics of complex and chronic conditions like the current stage of the disease, the rate of progression, or the ability of the patient to cope with and manage their medical condition [3-5]. Details such as these are crucial not only in planning the best course of treatment for the patient, but also for supporting the informational granularity often required for research and quality assurance. Chang found that ICD9 codes (structured data representing the diseases present in a patient) were often very specific (96-100%) compared to manual reviews of patient charts by clinical experts, but varied widely on how sensitive they were (28-70%) [6]. Most EMR systems are characterized by a wide array of structured data sources that are readily available and amenable to computing.

2.2 Use of unstructured patient documentation

The variety of ways that can be used to express information in clinical text means that although this data is rich and descriptive, much of the information is unavailable for use in research, decision support, and clinical decision-making. Often human beings must read the documentation or abstract data from the clinical records through manual chart review efforts. The goal of NLP is to provide automated means to assist and even replace portions of this manual process. NLP allows clinicians to continue to record pertinent information in a way that is natural (narrative text) and creates a formal representation of the information that computers can manipulate. In essence, NLP provides the means to extract information stored in clinical texts and makes it computable.

Researchers have used NLP to extract clinical concepts from the narrative portions of patient documentation. The addition of these concepts provides a much richer detail of the patient’s state of health. Hripcsak [7] and Sneiderman [8] have explored methods for identifying clinically relevant findings in narrative text and Meystre [9-11] has studied the ability of systems to find medical problems mentioned in clinical notes. More specific domains have been explored where concepts related to cancer, tuberculosis, heart failure, and even smoking status have been identified using NLP methods [12-15].

3. Providing structure to unstructured clinical text

The basic structure that NLP provides to unstructured text is called annotation. Annotations can be thought of as labels added to clinical text in order to identify key concepts or assign additional meaning. Any number of annotations can be made to text to add machine-computable information. For example, a sentence appearing in a radiology report such as ‘rule
out pneumonia.’ may be assigned annotations such as ‘possible’ and ‘pneumonia’ depending on the intended use case. Manual annotation is the human assignment of labels to spans of targeted text and is used to both train and evaluate the performance of NLP systems. Automated annotations are created from the output of NLP systems. Both types of annotation add structure to documents that can be used for secondary purposes - any process that can use structured data can benefit from text through the annotations.

Manual annotation is usually done at a level of granularity appropriate for a project-specific use case. This means that the same annotations may not be easily used for other projects. As manual annotation is labor-intensive and expensive, the inability to reuse annotations is an unfortunate waste of resources. If clinical documents were annotated with reuse in mind and stored in a standard, accessible way, new annotations could be added to documents with existing annotation to support more than just the initial use case.

A common approach to manually annotating patient documentation is having two clinical experts independently annotate a document and having a third adjudicate disagreements. This process itself introduces the need for effective methods of managing and analyzing annotation sets. In situations where annotation tasks are done on a large scale with many experts working in parallel on related tasks, management of data generated, reusability of annotations, and storage using markup that supports additional layers of annotation becomes even more important. These methods must be simple enough to implement for even small-scale efforts and flexible and robust enough to be used for large-scale clinical corpus annotation efforts.

3.1 The need for reusable annotations

One such large-scale annotation undertaking is underway in the Consortium for Healthcare Informatics Research (CHIR). CHIR is an informatics initiative of the Department of Veterans Affairs (VA) that concentrates on the development and application of methods that can effectively use unstructured data for clinical research and inference. These efforts encompass a multi-disciplinary program of research that includes experts in clinical domains, epidemiology, and informatics across many VA and VA-affiliated sites exploring NLP through targeted clinical use cases.

The VA is the largest health care system in the United States, providing care to approximately 6 million veterans at over 1,400 points of care. Each point of care feeds a common EMR system resulting in the generation of billions of transactions annually. One reason why incorporating the use of text in research and care has become so important in the VA, is that the majority of electronic clinical documentation is stored as text rather than as structured, coded data. As such, being able to access this data dramatically enriches the amount of information that is accessible for each patient across a longitudinal time frame.

Annotation efforts related to CHIR involve dozens of human annotators working on thousands of text documents to develop, test, and implement NLP systems that once operationalized, will be run on hundreds of millions of documents within the VA health care system. Over a four-year period, annotation associated with this consortium will be applied across varied document sources and clinical use cases.

All CHIR projects involve assessment of information extracted from clinical texts. CHIR projects will depend on manual annotation as well as annotation tasks that are assisted by various machine-aided approaches. Because these projects are highly inter-related and can build on each other, communication and planning are needed to ensure cooperation and generalizability across tasks. The integrated nature of the consortium depends on flexible sharing of output and tools between projects, rather than treating each project as a silo of activity. A core principle for these efforts is to draw from acceptable practices and incorporate existing standards to the extent possible. Achieving these goals relies on using a consistent approach for annotation efforts that support development of adequate reference standards specific for each project and provides a flexible framework for reuse.

3.2 Current annotation efforts

Previous efforts provided an operational model that shaped the base of CHIR annotation. While different models have been used to store annotations in the biomedical domain and in the NLP community at large, there are two common themes. First, many of the annotated corpora serve as a common reference standard that different researchers can use to compare their work with others as they attempt to improve upon state-of-the-art methods. These corpora are “completed works,” not meant to be added to or changed. For example, the Penn Treebank provides part-of-speech information on a few thousand documents and has become the standard for comparison when researchers attempt to improve the part-of-speech tagging functionality of NLP systems [16]. Similarly, the annotations from the Message Understanding Conferences (MUC) and Automatic Content Extraction programs (ACE) have provided a means of comparing
named entity recognition and coreference resolution systems [17, 18].

A second theme in existing annotated corpora is that annotations are stored in a way most convenient for the particular task. For example the Penn Treebank stores annotations in the form of bracketed tag sets. MUC annotations are stored in a mark-up language called SGML (a predecessor to XML). The GENIA corpus and other corpora available through the Linguistic Data Consortium use various schemata even when using consistent markup languages such as SGML or XML [19-20].

In the biomedical domain, a well-known annotation effort was undertaken by the Clinical E-science Framework (CLEF) corpus [21]. The CLEF corpus takes a whole patient approach, incorporating data from the entire patient record, and contains many layers of annotations including named entities, relations, modifiers, coreference, and temporal information stored in XML. While CLEF is one of the most extensive annotated health corpora, it has some limitations. For example, only a few hundred documents were annotated in this clinical corpus, annotation was done by only one team, and due to privacy and confidentiality issues these data have limited availability for other researchers to use.

Informatics for Integrating Biology and the Bedside (i2b2), a National Center for Biomedical Computing funded by the National Institutes of Health, has sponsored international competitions to develop solutions for particular NLP problems. While teams participating in these competitions develop solutions independent from each other, they all use the same corpus, and so performance results, system descriptions, and lessons learned can be collectively shared. The i2b2 challenges have provided valuable examples of methods that can be used for annotation. One of these was having members from each of the teams that built an NLP system to contribute manual annotations as well. In this way, the community built part of the reference standard for the 2009 challenge task. i2b2 has been known to employ rigorous annotation approaches across multiple challenge tasks [15, 22]. Possibly the most important contribution of the i2b2 corpora are that they de-identified and can be used by researchers after the competition in which they are used. For some years, the same documents have been annotated for different use cases. The reuse and availability of the i2b2 corpus makes this an exciting data source for all NLP researchers.

Chapman is leading an equally exciting effort to engage the community in building a repository of annotated clinical text [23]. This model also involves providing de-identified clinical text for use in NLP. One condition of gaining access to the records is that researchers will submit any annotations their teams make as part of their project back to the repository.

All of these efforts suggest that the goal of reuse of clinical corpora and storing annotations in a common, standardized way is achievable. The remaining challenge is operationalizing these efforts in a practical manner in which annotations can be used for multiple projects and clinical use cases. We propose such a method in our discussions that follow.

3.3 The CDA as an annotation storage method

The Clinical Document Architecture (CDA) is a document markup standard created by the Health Level Seven (HL7) standard development organization to facilitate the storage and exchange of structured documents [24]. The CDA standard has been used since 2000 in healthcare systems around the world and is becoming an important standard in the United States. It is used in many health exchange specifications developed by the American National Standards Institute (ANSI) and the Healthcare Information Technology Standards Panel (HITSP), and has been recognized for that purpose by the secretary of Health and Human Services in January of 2008 and again in January of 2009.

The CDA enables clinical content to be formally expressed using the HL7 Reference Information Model (RIM) – a detailed technical model for describing clinical elements and their attributes – and allows structured and unstructured content to be displayed together [25]. This allows coded data, or even external media, to be displayed along with narrative text.

Previous work has demonstrated that concepts, modifiers, and relationships can be identified in narrative text and stored in CDA documents when entered through an advanced EMR system using structured templates [26]. While these systems can take advantage of the CDA as new documents are created, immediate benefits can also be gained by using legacy data sources. Narrative text can be inserted as the payload of a CDA document without modification. As further structure and coding are introduced, they can be directly incorporated into the document. The more detail and structure that is added, the more potential the document has for reuse. The CDA stores annotations inline with the text, so even annotations stored in standoff files separate from the original document require only a simple transformation [27].

The CDA has the ability to natively represent important information that may be gathered in the annotation process including document meta-data, concepts with modifiers and relationships, and mappings of concepts to coded terms. CDA documents record clinical information meant to persist and have
clinical and legal stewardship. Because of this, the standard specifies structure for recording revisions and changes to the original content. In the case of annotation, changes to the original text and subsequent changes to annotation can be recorded and assigned authorship, clearly delineating content produced by the original clinician and that produced through annotation and additional layers of review.

Storing annotations in CDA documents affords the benefit of reusability of legacy data, especially in facilitating annotation tasks that build upon each other. A variety of tools exist that use the CDA for decision support, display, and other EMR functionality [28-31]. Having the annotations as part of the actual document, allows these tools to use annotations without further development.

4. Annotation task description

The first step in any annotation task is the development of an annotation model that defines which concepts are marked and the way in which they should be marked. Much work has been done to define generalizable annotation models [21, 32]. These models are similar whether one is discussing them in terms of human or machine tasks and each task can inform the other. To store annotations in the CDA, an annotation model must be mapped to classes of the RIM. Most annotation concepts fall under the RIM Act class, which includes orders, observations, substance administrations, supplies, and clinical documents. For a comprehensive list of RIM capabilities, the HL7 RIM specifications can be consulted [33].

4.1 Processing text in steps

NLP operates in a pipeline-fashion, in which each process can build upon the output from previous processes along the pipeline. Steps later in the pipeline often perform higher order tasks that depend on more basic previous steps. For example, the assignment of parts-of-speech (e.g., noun, verb) is dependent on the results of a component that breaks text into sentences and individual words.

Parallel to the NLP pipeline, annotation can be divided into tasks that identify simple structure, detect clinical concepts and relationships, and make clinical inferences. These steps of manual annotation inform NLP development tasks. Table 1 presents the different types of annotations commonly used in NLP. These layers of annotation are based on linguistic structures and in our experience, represent a comprehensive set of possible annotations [34]. Each annotation level and associated task is presented below with some example representations in CDA format. To illustrate the potential of the CDA to support annotation, we present a sample patient with chronic asthma visiting a specialty clinic for an episode of an acute exacerbation. The entire CDA document for this example is available from the authors.

<table>
<thead>
<tr>
<th>Annotation Level</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta</td>
<td>Structure</td>
</tr>
<tr>
<td>Document</td>
<td>Structure</td>
</tr>
<tr>
<td>Lexical</td>
<td>Structure</td>
</tr>
<tr>
<td>Syntax</td>
<td>Structure</td>
</tr>
<tr>
<td>Semantics</td>
<td>Concepts</td>
</tr>
<tr>
<td>Discourse</td>
<td>Concepts and Relationships</td>
</tr>
<tr>
<td>Pragmatics</td>
<td>Inference</td>
</tr>
</tbody>
</table>

Table 1. Levels of annotation with associated tasks

*Meta annotations* are information about a document as a whole and include data such as document type, originating institution, patient, and clinician who authored and signed the document. The CDA provides structure for meta annotations in its header – the portion of a CDA document that informs discovery, management, and retrieval. *Document annotations* include information about structure – which template was used, identification of sections and section headers, and identification of paragraphs. CDA inherently provides representation for document annotations.

*Lexical annotations* are analogous to document annotations and describe structure within paragraphs such as token and sentence boundaries. *Syntax annotations* usually include part-of-speech information for single words and grouping of phrases across multiple words. Syntax and lexical annotations are almost always calculated, but often are not included in the final annotation schema because they are rarely an end in themselves; rather a means to identifying more important, higher-level annotation.

*Semantic annotations* are some of the most common in NLP. One reason for this is that they are relatively easy to find and are the first level of annotation to approach the “structure from unstructured” benefit of NLP methods applied to clinical text. Semantic annotations identify concepts and concept attributes. The CDA can support both simple and compound concepts, allows mapping to controlled terminologies or coding systems, and includes a range of features relevant to semantic annotations such as negation, severity, temporality, experiencer, certainty, change, and anatomical location. Figure 1 shows the concept “asthma” annotated along with the modifiers severity:
The patient has had moderately severe asthma since he was ten that has become worse since his last visit.

Discourse annotations describe the relationship between semantic annotations. These include concept-value pairs and the relationships between concepts such as a certain condition and a test that is ordered or a medication that is prescribed. Figure 2 shows a concept-value pair of a pulmonary function test called spirometry, the result of the test: less than 70%, and that the indication of the test was the patient’s asthma: a reference to the first observation (OBS1), which is shown in Figure 1 to be asthma. The RIM provides many types of relationships that can be described in the CDA including causality and indication.

Pragmatic annotations rely on clinical inference. They are different from all other annotation levels, because instead of just adding structure, pragmatic annotations add clinical content. One of the most common types of pragmatic annotation classifies a patient as belonging to a specific cohort based on the information within a document. To do so, NLP systems look for semantic and discourse annotation that provide evidence for making a clinical inference. Figure 3 shows a pragmatic annotation that concludes the patient fits the definition of having particular type of asthma based on evidence found in annotations in the note.

4.2 Use cases for annotation reuse

From these levels of annotation, it is possible to envision a scenario where annotation tasks occur in several phases, each possibly being done by different domain experts as components of different but possibly related projects. The first phase could be the use of a tool that identifies concepts in the Unified Medical Language System (UMLS). Several tools have shown to be proficient at this task [35]. Next, human annotators or a more complex system could identify attributes present in the text that modify the concepts. Another phase of annotation could include connecting concept and value pairs. Another could identify relationships between concepts. A final set of phases could be used to make relevant clinical inferences using all available annotations.

This phased approach fits the current model where annotation is done for specific purposes and when captured in the CDA, lends itself well for reuse. With each new use case, previous annotations can be built upon and groundwork can be laid for future annotation tasks. The end result is a corpus of clinical notes that has been reviewed many times for various domains and focused tasks.

Using this approach, documents like the consultation note for our sample patient may eventually end up containing annotations for all conditions, medications, and procedures the patient has had, even those unrelated to asthma. Because the annotations are stored together with the clinical text, the annotations relevant to asthma can still be analyzed for our project and they can be used all together for
coding and billing, identifying research cohorts, or making other clinical inferences utilizing additional layers of annotation.

5. Discussion

This paper explores the use of CDA for storing annotations created through manual annotation and NLP. Operationalizing the way annotations are stored and how they are collected in order to facilitate reuse is of great importance when limited resources must be leveraged and multiple tasks must be accomplished.

Still, some differences exist between clinical documentation (the purpose of the CDA standard) and the annotation process that merit discussion.

5.1 Preserving original clinical content

Although the CDA was designed to support patient care, it is a logical home for annotation, because the annotation process in essence identifies information that describes patient care. As the use of NLP methods in medicine continues to mature, annotation will play an increasingly important role in supporting patient care. In the meantime, storing...
annotations in a CDA document means adding to the original clinical content. The CDA provides structure for separating additions and modifications that occur at a later date in a few different ways. First, a CDA document can be marked as a transform of an original document and can include a reference to the original, unannotated note. Second, authorship can be assigned to different parts of a document allowing content provided by the original author to be displayed alone. Finally, the CDA has revision tags that can allow original content to be displayed along with modified content.

5.2 Assigning annotation authorship

An interesting issue that arises in annotation for NLP is that the annotator is not always a credentialed health care provider or even a human. In the CDA, the Author role is defined as somebody or something that creates information from a set of knowledge or skills. Under this definition, systems that create clinical content can be authors. Another role already defined in the standard that may be used is that of Data Entry. We suggest using the Author role when adding inferences and clinical content that requires medical judgment and the Data Entry role when marking structure on existing clinical content.

5.3 Multiple document renderings

CDA documents are generally displayed through the application of Extensible Stylesheet Language Transformations (XSLT), which is used to transform XML into HTML or PDF. As such, the same CDA document can be displayed with just the original text, highlighting specific concepts, or showing all annotations by using different style sheets. This allows all annotations to be stored in the CDA with only those that are relevant displayed. In this way, not only can annotated corpora take advantage of tools built for the CDA, but tools used to display and transform XML as well.

5.4 The place of the CDA in the EMR

Although we started the paper talking about patient documentation in the EMR in general, our emphasis is the storage of annotation in the clinical text portions. The overall architecture of EMR systems, the storage of structured data, and representing data longitudinally across episodes of care are active areas of research, but outside the scope of this work.

6. Conclusions and Future Work

As the number of proposed secondary uses of data extracted from the EMR increases, so too will the importance of standards that promote more efficient information reuse. In an environment where resources are limited and review tasks are costly and have high workload, the combined goals of reuse and layered annotated data are of great importance. This paper describes a starting point, the decision to annotate with reuse in mind and applying the CDA as a scalable method to store multiple layers of annotations with the clinical texts.

While part of ongoing research where a fully specified CDA annotation model will be implemented and integrated with our current annotation tools, this study successfully completes some key first steps: We have shown that the CDA represents an existing standard amenable to supporting both human and machine annotation tasks and shown how annotations map to CDA elements. Some levels of annotation, such as lexical and syntax annotations, may be too verbose and not appropriate for inclusion in the long-term

| Figure 3: A pragmatic annotation inferring the patient should be classified in the cohort: Intrinsic asthma, with (acute) exacerbation based on evidence of asthma, spirometry ratio and albuterol inhaler. |
storage of a document. As these annotations can be readily calculated, and are often only used in the process of creating higher-level annotations, this may not be too much of an issue. Further work will determine the appropriate placement for all levels of annotations produced for a document.

The adoption of the CDA as a vehicle for annotation storage has the potential to promote the reuse of annotations, reduce the cost of conducting research in the areas of clinical NLP, and increase the value of legacy data resources.

7. Acknowledgements

This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding support from the VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204; and the Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374.

8. References


