Health Level Seven Interoperability Strategy: Big Data, Incrementally Structured

R. H. Dolin1; B. Rogers2; C. Jaffe3

1Orange, California, USA;
2Apixo, Inc, Advanced Development Department, San Mateo, California, USA;
3Health Level Seven International, Del Mar, California, USA

Keywords
Big data, CDA, HL7, Semantic Interoperability

Summary

Objectives: Describe how the HL7 Clinical Document Architecture (CDA), a foundational standard in US Meaningful Use, contributes to a “big data, incrementally structured” interoperability strategy, whereby data structured incrementally gets large amounts of data flowing faster. We present cases showing how this approach is leveraged for big data analysis.

Methods: To support the assertion that semi-structured narrative in CDA format can be a useful adjunct in an overall big data analytic approach, we present two case studies. The first assesses an organization’s ability to generate clinical quality reports using coded data alone vs. coded data supplemented by CDA narrative. The second leverages CDA to construct a network model for referral management, from which additional observations can be gleaned.

Results: The first case shows that coded data supplemented by CDA narrative resulted in significant variances in calculated performance scores. In the second case, we found that the constructed network model enables the identification of differences in patient characteristics among different referral work flows.

Discussion: The CDA approach goes after data indirectly, by focusing first on the flow of narrative, which is then incrementally structured. A quantitative assessment of whether this approach will lead to a greater flow of data and ultimately a greater flow of structured data vs. other approaches is planned as a future exercise.

Conclusion: Along with growing adoption of CDA, we are now seeing the big data community explore the standard, particularly given its potential to supply analytic engines with volumes of data previously not possible. In this report, we will describe case studies showing how CDA has been leveraged by a big data analytic engine to glean new insights that will potentially lead to care improvements.

1. Introduction

Health Level Seven International’s (HL7) Clinical Document Architecture, Release Two (CDA), a foundational standard under the United States Meaningful Use (MU) program, contributes to a “big data, incrementally structured” interoperability strategy, whereby data structured incrementally gets large amounts of data flowing faster. Prior articles have described CDA[1]. Here, we extend the discussion, to describe how CDA supports the early exchange of large volumes of loosely structured documents, and sets a glide path for the incremental evolution of greater structure.

The CDA strategy for incremental interoperability was defined close to fifteen years ago, in part to meet the needs of front line clinicians to share narrative documents balanced against the enterprise need for structured data. We are now seeing the big data community explore CDA, particularly given its potential to supply analytic engines with volumes of healthcare data previously not possible. In this report, we will describe case studies showing how CDA has been leveraged by a big data analytic engine to glean new insights that will potentially lead to care improvements.

2. Background and Significance

2.1 Background on HL7

For nearly three decades, HL7 (http://www.hl7.org/) has been the world’s leading developer of healthcare data interoperability standards. Its portfolio of products and services have evolved from simple messaging paradigms to a broad range of specifications that support patient care, public health, basic and clinical research, and data analytics.

Since HL7 first received accreditation from the American National Standards Institute (ANSI) in 1996, it has won adoption by national standards bodies worldwide, including the International Organization for Standardization Technical Committee for Health Informatics (ISO TC 215). In fact, it is the authorized or de facto standard in more than 30 countries. The first standard published by ISO was the HL7 version 3 Reference Information Model (RIM), from which many other specifications are derived. Today, many
HL7 standards are balloted within ISO and become ISO standards as well.

During the last decade, HL7 has grown in its adoption through its collaboration with other standards development organizations, including International Health Terminology Standards Development Organization (IHTSDO), which develops SNOMED CT; Regenstrief Institute, which develops LOINC; Workgroup for Electronic Data Interchange (WEDI), the standards of which enable financial transactions; Integrating the Healthcare Enterprise (IHE), creating profiles for specific interoperability paradigms; Clinical Data Interchange Standard for Exchange, Transformation and Storage (CDISC), that provides data interchange standards; and the Joint Commission on Accreditation of Healthcare Organizations (JCAHO), which has created the SCRIPT standard for electronic prescribing; Digital Imaging and Communications in Medicine (DICOM), which has developed standards for handling, storing, retrieving, archiving and transmitting radiographs and other medical images; and American Society for Testing and Materials (ASTM), which is the international standards organization responsible for the development of the Continuity of Care Record and which joined with HL7 in the creation of the Continuity of Care Document standard.

These partnerships have facilitated ever increasing implementation strategies and innovative approaches to data management. In 2008, HL7 began the Joint Initiative Group (JIC) with ISO and the European Committee for Standardization (CEN), in order to improve the creation of better standards among global standards development organizations.

In addition, HL7 specifications have grown to include a model framework for electronic health records, for services-aware architecture, and for various decision support paradigms. These standards have been enhanced by the application of specific profiles that constrain the standards for individual communities of users (e.g., Cardiology and Pediatrics) and for specific functionality (e.g., continuity of care, electronic prescribing, decision support, and public health reporting).

### 2.2 Background on CDA

CDA became an ANSI-approved HL7 Standard in May 2005 [1,2]. CDA is a document markup standard that specifies the structure and semantics of a clinical document (such as a discharge summary, progress note, procedure report) for the purpose of exchange. A CDA document is a defined and complete information object that can include text, images, sounds, and other multimedia content. It can be transferred within a message, and can exist independently, outside the transferring message.

CDA documents are encoded in Extensible Markup Language (XML). They derive their machine processable meaning from the HL7 Reference Information Model (RIM) coupled with standard terminologies. The RIM provides a powerful mechanism for enabling CDAs incorporation of concepts from standard coding systems such as SNOMED CT and LOINC.

The CDA model is richly expressive, enabling the formal representation of clinical statements (such as observations, medication administrations, and adverse events) in order for that representation to be interpreted and acted upon by a computer. On the other hand, CDA offers a low bar for adoption, providing a mechanism for simply creating a document with a structured header and sections containing only narrative content. The intent is to facilitate widespread adoption, while providing a path for incremental semantic interoperability.

The CDA specification is richly expressive and flexible and is designed to be broad enough to include the entire domain of clinical documents. Templates and/or implementation guides can be used to constrain the CDA specification for a particular use and to provide validating rule sets that check conformance to these constraints.

#### 2.2.1 CDA’s Role in US Meaningful Use

The US Health Information Technology for Economic and Clinical Health Act of 2009 (HITECH) provides the US Department of Health & Human Services (HHS) with the authority to establish programs to improve health care quality, safety, and efficiency through the promotion of health information technology, including electronic health records and private and secure electronic health information exchange. Under HITECH, eligible health care professionals and hospitals can qualify for Medicare and Medicaid incentive payments when they adopt “certified EHR technology” and use it to achieve specified objectives.

Certified EHR technology is defined by the HHS Office of the National Coordinator for Health Information Technology (ONC) through a series of “Meaningful Use” regulations. Stage 1 of Meaningful Use (MU1) established criteria for standardized data capture and data sharing. Today, Stage 2 of Meaningful Use (MU2) extends the criteria for certified EHR technology by raising the bar on required interoperability standards, including standards designed to support transitions of care and clinical quality reporting[3]. In 2016, Stage 3 of Meaningful Use (MU3) will build upon MU2, tying our ability to measure care with interventions that improve clinical outcomes.

As noted above, implementation guides can be used to constrain the CDA specification for a given scenario and to provide validating rule sets that check conformance to these constraints. Leveraging this capability, MU2 cites a number of CDA-based standards (Table 1).

The past two years has witnessed the growing global adoption of CDA-based standards. In addition, in the US there has been a dramatic increase in the number of CDA-based standards cited under MU2 vs. MU1. This can be attributed to several factors, but perhaps the most important are

1) **Incrementalism**: CDA is built on a principle of incrementalism, whereby there is a relatively low bar to creating a minimally conformant CDA that meets the human readability needs of front line clinicians today, but can be incrementally structured
Table 1  CDA content exchange standards under Meaningful Use Stage 2

| § 170.205(h) | CDA Guide for Quality Reporting Document Architecture, Category I (QRDA-I) [5]: Standardized representation of quality data for an individual patient. Data in a QRDA-I report can be consumed by a calculation engine to determine if the patient met the numerator or denominator criteria for a given quality measure. |
| § 170.205(i) | CDA Guide for Reporting to Central Cancer Registries [6]: Standardized cancer registry reporting format. |
| § 170.205(k) | CDA Guide for Quality Reporting Document Architecture, Category III (QRDA-III) [7]: Standardized representation of aggregate quality data (e.g. number of patients meeting the numerator criteria for a given quality measure). |

There are many different kinds of templates that might be created. Among them, the most common are document-level templates (e.g. Consult Note Document Template, Progress Note Document Template) that constrain fields in the CDA header, and define containment relationships to CDA sections; section-level templates (e.g. Physical Exam Section Template, History of Present Illness Section Template) that constrain fields in the CDA section, and define containment relationships to CDA entries; and entry-level templates (e.g. Blood Pressure Entry Template, Discharge Diagnosis Entry Template, Functional Status Entry Template) that constrain the CDA clinical statement model in accordance with real world observations and acts.

A key point is that human readable narrative content is always required in a CDA document. Structured entries, governed by CDA templates, serve to consistently formalize some or all the clinical statements present in the body of the document. Thus, with CDA it is relatively easy to get narrative documents flowing. Moreover, CDA provides a path for adding more formal structure through the application of templates.

2.2.2 HL7’s “Big Data, Incrementally Structured” Approach

According to the Healthcare Information and Management Systems Society (HIMSS) Health Story Project [8], more than a billion clinical documents are produced in the United States each year. These documents comprise around 60% of clinical information, the majority of physician-attested information, and are used as the primary source of information for reimbursement and proof of service. Health Story project members believe that this tremendous source of clinical information is underutilized in current computer-based record systems, and together with HL7 have developed the incremental interoperability approach illustrated in Figure 1. In this approach, clinical documents, in various formats, can, with relatively low effort, be migrated into a semi-structured minimally conformant CDA format, enabling the exchange of large volumes of narrative (and partially encoded) data that carry consistent metadata and that can be indexed for efficient communication and retrieval. From there, we can incrementally layer on prioritized CDA templates. With this approach, we gradually attain progressively more structured content to drive quality reporting and decision support, while continuing to meet the needs of front-line clinicians who simply need to read the consultant’s note or to read the recent discharge summary.

Incrementalism refers not only to the ability of the HL7 standards community to incrementally structure narrative documents, but also to the recipient who can incrementally glean new ways of processing and analyzing the content. This latter point is worth emphasizing – incrementalism works both on the side of the sender AND on the side of the receiver. Much evidence indicates that the rich narrative conveyed in a clinical document can be structured after the fact by the recipient [9, 10]. Whereas a classic relational database approach to data warehousing and analytics will support answering questions that fit into the database schema, a big data analytic approach supports posing and attempting to answer new questions. The raw source data, that is the rich narrative, is present regardless, and big data analytic engines are incrementally discovering new ways to harness that data.
of a network model for a care delivery system based on structured and unstructured data, and investigates the computational power afforded by such a network model.

All clinical data for both cases were collected from EHR systems of healthcare organizations that have more than 75% of their member physicians using an EHR as their primary method of clinical documentation. These organizations use a specific analytics platform to provide physicians and case managers with a single point of access to system-wide clinical data for each patient. Research studies on de-identified data sets extracted from this platform are allowed under the terms of use of the platform, and do not require specific IRB approval.

Clinical data for the first study was extracted from an Independent Physician’s Association (IPA) with approximately 65,000 patients. The data spans the entire spectrum of the ambulatory care experience for these patients, including nearly all specialties. The results reported below apply to a small cohort of patients randomly selected from the entire population.

Clinical data for the second study was extracted from a healthcare system that includes multiple specialties and a teaching hospital. The results reported below apply to a cohort of 1,752 patients randomly selected from the entire population. In this dataset, as much as 10 years of historical EHR data were available for patients in some practices and the average history for patients was over three years.

The analyses for both case studies were enabled by the distributed computing architecture of the big data analytics platform that allows it to aggregate and analyze large volumes of clinical data simultaneously, using the following steps:

1. All clinical data were extracted from the EHRs, including structured data, narrative data such as encounter notes and administrative documents, and images, including scanned discharge summaries, labs, diagnostic reports and consult letters. All data was uploaded securely to the cloud-based analytics platform.

2. Uploaded data was then processed on a Hadoop-based distributed computing platform that automatically identified steps required to map the data to a normalized patient object representation. The normalized patient object is a mathematical representation of the entire known clinical history of a patient, including encounters, measurements, observations, communications and other events that are relevant to the patient’s health, viewed as a sparse vector in a very high dimensional feature space, where patient clinical attributes are the features. The sparse vector representation is constructed as follows: Structured data, including attributes such as measured value, timestamp, mappings to known coding systems and provenance of the data are recorded in the patient object. Unstructured portions of each document in the patient record, including text extracted from images, are tagged with clinical concepts from an application-specific dictionary using text mining techniques. Values relevant to these concepts are extracted from text. For example, a progress note containing the text “Urinalysis, dipstick...Protein: 2+” might be tagged with the LOINC code 2888-6 and the extracted value “2+.” An example of a similar text mining procedure for clinical concept extraction is described by LePendu et al. [11].

3. Data for each case study is extracted by a map-reduce job in Hadoop, resulting in the various “event streams” described below.

### 3.1 Case Study: Inadequacy of Coded Data

Gandhi et al. [12] showed that even in a highly sophisticated healthcare system such as Partners HealthCare System and Harvard Medical School, key clinical information is missing from the coded layer of the EHR. Specifically, they found that for patients who had undergone a splenectomy, 71% of the time this was documented only in clinical narrative and was absent from the coded problem list. In an independent study of 20,000 patients randomly selected from the enterprise EHR of an IPA, we have found that 65% of key clinical information representing a broad spectrum of conditions is missing from the coded layer of any single EHR system [13].

Here, we constructed an event stream by searching clinical documents for patterns that indicated the presence of evidence for membership in the quality measure numerator or denominator. For NQF 0619: Diabetes with Hypertension or Proteinuria, this included evidence for Diabetes Mellitus, for a measurement of proteinuria, etc. For NQF 0067: Coronary Artery Disease (CAD): Oral Antiplatelet Therapy Pre-
scribed for Patients this included evidence for CAD, for antiplatelet therapy, etc.

For each quality measure, the event stream was queried for evidence that each patient should be included in the numerator and denominator and a comparison of the computed measure values based on coded data alone and coded data plus unstructured data was made.

3.2 Case Study: Network Model for Referral and Care Management

Here we demonstrate the construction of a referral network that represents patterns of care in a large health system. The referral network is a weighted graph whose nodes represent care providers or sites of care, and whose directed edges represent referral paths. The weights on these edges can represent any number of characteristics derived from the population of patients who have been referred along the associated edge.

To construct such a referral network from structured data alone is virtually impossible (provider compliance with referral tracking workflows in the EHR systems is low, patients often seek the recommended care from a provider other than the one referred to, etc.). Our approach was to extract outgoing referrals and the care events that result from them from clinical narrative in CDA documents.

In practice, this is equivalent to the construction of an event stream which identifies evidence in documents for referral activity. Here, we first constructed an event stream that identifies key terms indicating an increased likelihood that a document contains evidence of a referral. This event stream was then queried to generate a second event stream of documents to search for referral element details in a separate map phase.

We generated the search patterns for the second event stream using a machine learning model that automatically generated a large number of candidate documentation patterns that were associated with outgoing and incoming referral activities. The parameters of the model were optimized over a very large number of documents, using the small fraction of referrals that were known from structured data to guide the optimization. The critical value of CDA in this context is that it preserves clinical narrative from many different care contexts, along with whatever structured data is available, so that it can be applied to analyses that were not anticipated at the time the data was generated and stored, but which are nevertheless highly valuable. Furthermore, because the structured data that is relevant to the document in question is encoded along with the narrative, it is possible to compute a vast number of different edge weights for referral networks that are constructed using CDA documents. Each set of distinct edge weights can be used for a specific application, for example to identify sections of the network that lack a particular resource, or to identify referral patterns that result in higher costs and worse patient outcomes.

To illustrate the power of this approach, we constructed a small Care Network and then performed a query over this network for three chronic diseases (liver disease, chronic kidney disease, and diabetes) to look for measurable differences in patient attributes when computed over different parts of the Care Network. Figure 2 depicts a subset of the care network for liver disease patients. Dark grey rectangular nodes are consulting providers; light grey oval nodes are referring providers. The main cluster is centered on referrals to GI clinic. The cluster on the left is centered on ED visits. We further identified a subset of these patients who had at least one specialist consultation or ED visit for which the patient was noted as having no referring physician (“self-referral cohort”). Using the coded data associated with these referral events, we then computed for both the self-referral cohort and its complement in the base cohort, the average ALT value for liver disease patients, Hemoglobin A1C for diabetic patients and serum creatinine for chronic kidney disease patients.

4. Results

In 1959 Mantel and Haenszel described the potential for bias and statistical error in retrospective analyses of observational data[14]. Here, we are reporting data
mining results based on the methods described above, and would note that the findings have not been subjected to confirmation in prospective studies in real clinical settings.

### 4.1 Case Study: Inadequacy of Coded Data

For **NQF 0619**, compliance increased from 22 percent to 45 percent when clinical narrative was included in the analysis (Figure 3). For **NQF 0067**, compliance moved from 37 percent to 44 percent with the inclusion of clinical text (Figure 4). What is perhaps more significant for the NQF 0067 result is that the cohort of patients eligible for the measure changed dramatically when information extracted from clinical narrative was included in the measure computation, increasing from 71 to 106. Furthermore, a significant number of additional exclusions were identified in the textual data, resulting in a different final cohort of patients who were eligible for the measure. We conclude that coded data is not sufficient to accurately identify patients for inclusion in quality measures, or to track their compliance, and that information contained in clinical narrative enables more accurate quality measurement.

### 4.2 Case Study: Network Model for Referral and Care Management

We found that for liver disease patients there is a statistically significant increase of 11.7 in the ALT measured for patients with a self-referral identified within their care paths, relative to patients with no self-referrals identified in the patient history. Small differences in measured values of serum creatinine for kidney disease patients, and Hemoglobin A1c for diabetic patients were not statistically significant. These results are shown in Table 2.

We conclude that noise levels in data extracted from unstructured data using this technique are low enough to make statistically meaningful measurements of patient clinical characteristics. With further analysis it could be possible to draw conclusions about the clinical relevance of these findings. Even at this level of detail it is significant that we have identified a difference in coded results between patient cohorts based solely on attributes of their paths of care through the health system.

### 5. Discussion

While few would dispute that clinically relevant data resides in narrative documents, an outstanding question from a national interoperability perspective is whether or
not shifting focus to get large volumes of semi-structured narrative documents flowing will facilitate getting us to the types of structured data elements needed to drive decision support, quality reporting, and other needs. In healthcare we are challenged by seemingly competing needs – the clinician’s need for fast and intuitive data capture, the enterprise need for structured and coded data, the clinician’s need to minimize disruption to workflow, the software developer’s need for unambiguous data.

The CDA “big data, incrementally structured” approach in some ways can be considered going after discrete data elements INDIRECTLY, by focusing first on the flow of narrative, which is formatted per a standard that supports the incremental addition of greater structure. And while there is growing evidence that big data analytics can harvest CDAs effectively, the question remains – is the CDA approach the most effective way to get us to the structured and coded data needed to drive improvements in healthcare?

In Figure 5 we present a set of qualitative and unsubstantiated graphs, thought exercises if you will, comparing the CDA approach to an approach based on the communication of discrete data elements such as might be seen in an HL7 Version 2 message. Graphs assess overall data volume (which will be greater with the CDA approach), structured data volume (which will be smaller with the CDA approach initially, but where the CDA approach will increase more rapidly, ultimately overcoming the discrete data approach), and overall implementation cost (where we assume a lower cost per data element with CDA). As a future exercise, we plan to study these graphs, gathering quantitative observations in order to determine whether or not they are realistic.

### 6. Conclusion

CDA, a foundational standard in MU, contributes to a “big data, incrementally structured” approach. The next version of Consolidated CDA, due to be published Summer 2014, will enhance the standard with the addition of new document types (Transfer Summary, Referral Note, Care Plan), and additional templates that support a growing number of use cases for transitions of care, decision support, and clinical quality measurement. At its most basic level, Consolidated CDA will continue to strive to meet the narrative interoperability needs of clinicians, while incrementally marching forward with additional structured content.

### References

3. Department of Health and Human Services, Office of the National Coordinator for Health Information Technology. Health Information Technology: Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, 2014 Edition; Revisions to the Permanent Certification Program for Health Informa-