Veterans Health Administration Experience with Data Quality Surveillance of Continuity of Care Documents: Interoperability Challenges for eHealth Exchange Participants

Jay Lyle, PhD, Omar Bouhaddou, PhD, Nathan Botts, PhD, Marie Swall, Eric Pan, MD, Terry Cullen, MD, Margaret Donahue, MD, Nelson Hsing, ScD*

US Department of Veterans Affairs, Washington, DC

Abstract

As part of ongoing data quality efforts authors monitored health information retrieved through the United States Department of Veterans Affairs’ (VA) Virtual Lifetime Electronic Record (VLER) Health operation. Health data exchanged through the eHealth Exchange (managed by Healtheway, Inc.) between VA and external care providers was evaluated in order to test methods of data quality surveillance and to identify key quality concerns. Testing evaluated transition of care data from 20 VLER Health partners. Findings indicated operational monitoring discovers issues not addressed during onboarding testing, that many issues result from specification ambiguity, and that many issues require human review. We make recommendations to address these issues, specifically to embed automated testing tools within information exchange transactions and to continuously monitor and improve data quality, which will facilitate adoption and use.

*Correspondence to: Nelson Hsing, Department of Veterans Affairs, Office of Interoperability (Nelson.Hsing@va.gov)

Introduction

The United States Veterans Health Administration (VHA) provides healthcare for almost nine million veterans. Increasingly, this care is provided in coordination with private sector providers. When care responsibilities are distributed among different providers, it is critical that all participants stay informed of each other’s activities and decisions so that care continuity can be maintained for the patient. Reliable quality data is needed for referrals and transitions of care: this is a key driver development of the VA’s Lifetime Electronic Record (VLER) Health program. The Healtheway initiative, a public-private collaborative that promotes standardized and trusted exchange of health data among participating member organizations, provides the means through which the VA could effectively test this type of national-level health information exchange.

In 2012, the VHA participated in a pilot program to test the use of Healtheway’s eHealth Exchange (at the time known as the Nationwide Health Information Exchange, or NwHIN). This pilot program included VA, Department of Defense, and 12 extant Health Information Exchanges (HIEs) as part of the VA’s VLER program. Subsequently, in 2013 the project became operational, and now includes upwards of 45 exchange partners with whomover 80,000 Continuity of Care Documents (CCDs) have been shared. VA is currently exchanging CCDs based on the HITSP C32 and C62 formats, and has initiated the process of enhancing the infrastructure to incrementally migrate partners to the Consolidated CDA document specification (C-CDA).

Research previously conducted by the VLER team, based on the VLER Health pilot period, identified several important data quality findings, including issues with C32 conformance to the specifications, optionality in the specification that adds to interoperability problems, and evidence that current testing tools cannot effectively validate clinical data during clinical operation. D’Amore, et al. showed that data quality issues persist even after migration to the newer C-CDA standard and after EHR systems have been certified based on the Meaningful Use criteria. At the same time, we showed that even pilot phase data monitoring provided critical enhancements to the data quality. These findings suggested that ongoing data quality monitoring and quality surveillance would continue to enhance adoption, decision-making and the quality of care for veterans using multiple providers by continuing to identify potential areas of miscommunication and interoperability challenges.
Comparison of VA data quality evaluation results to those gathered by other researchers in this domain has identified similar classes of issues independent of the specific study settings (3). In this paper, with over a year of experience in production, we report specific data quality issues presented by C32 documents provided to the VA. These have informed a data quality surveillance framework that includes the use of automated tools and systematic scoring of received documents.

**Objectives**

We posit that there are two key goals for improvement of the current quality of exchanged data: 1) reliable automated data integration that supports semantic interoperability and 2) increased clarity and usability of the data when accessed by care staff. In the long term, health systems should support automated semantic interoperability, including rule-based alerts and automated integration of received data into the health record. It is widely recognized that interoperability and decision support standards are not yet mature enough to support large-scale implementation of this sort of functionality (incremental improvement to support such a goal being a motivator for this research). However, while we are not yet able to completely rely on automated interfaces for the management of clinical data, current health information exchange standards increasingly support consistent representation of health record data to human reviewers. Although the current operation does not automatically incorporate received data in the patient health record, it does present externally sourced documents to clinicians on demand. The near-term requirement, then, is that the data be comprehensible to the viewing clinicians. It is important to note that clinicians may review the data not only one source document at a time, but more naturally, as combined data aggregated from multiple source documents. The first option can be satisfied by the narrative block of the CCD, but the second option requires parsing of the structured entries, and thus requires that these structured entries share design assumptions to a degree approaching that required for automated integration.

Whether a document meets one of these two requirements for a given element does not imply that it meets the other: a document may be structured correctly but provide information that is confusing to read, or it may provide clearly comprehensible information in a file that fails to conform to the rules defined by the standard specification. The focus of this research is to identify cases where information may be confusing or incomplete to a human viewer so that the specification may be refined—ideally to the point where it may support even more automated functionality—and to identify opportunities to leverage automated standards conformance tools to assist in that effort.

**Methods**

We have identified several cases where incoming documents do not conform to the published specification. We provide these cases to the institutions from which they were received as feedback to support their document generation improvement efforts. We score these documents according to their richness (how many of the sections supported by C32 are sent for a given sample of patient cases), quality (percentage of present fields filled in correctly), and semantic interoperability (percentage of fields of coded data type that are provided in the specified standard terminology). We are in the process of automating this process, as it is extremely time-consuming to perform manually. The automation will also allow us to process a larger sample of documents, perhaps even every document received, and achieve more accurate scores that account for data quality variability from one patient record to another.

Of equal interest are documents where a field instance may conform to a specification but fail to provide unambiguous information. We also find that assessing these issues with an eye to automation facilitates unambiguous definition of quality criteria even for manual review.

We designed a two-stage process to address both questions. In the first stage, a document is tested for conformance to specification by automated tooling; in the second, the document is inspected for clarity in a human interface. This process identifies a majority of technical errors, and it also identifies clarity issues not detected by automated tools. Further, by conducting the automated tests first, we would pre-populate the manual scoring template, reducing the effort required and allowing the analyst to focus on otherwise undetectable semantic issues.

Figure 1 shows the process at a high level.
Referring to the numbering in Figure 1, incoming documents (1) are automatically scored by existing tools, such as Model Driven Health Tools (MDHT) (2), or by other products, such as the SMART CCDA evaluation tool (3). The results of these automated activities are output in a common format (4) and stored in a database (5). The database supports an analyst screen (6) for the document, in which an analyst can identify problems found while inspecting the document in the clinician’s VistaWeb interface (i.e., VA clinician user interface) (7); this input is also stored in the database (5). The resulting detailed database can then support any number of reports, scoring algorithms, or other analytical processes (8). These processes can even employ different standards of conformance; e.g., whether and when a particular conditional field should be considered required.

We encountered a number of issues in this effort, including most notably a shortfall in tooling functionality and difficulty in distinguishing the causes of clarity issues.

XML parsers can provide validation to the CDA schema. For example, the National Institute of Standards and Technology (NIST) CDA validation tools perform much of the validation to the template constraints; however, it
provides only limited support for assessing conformance to terminology constraints. And, while the validation that it provides is useful guidance for developers, it is not designed for monitoring quality in a production environment by detecting or counting issues over time, across multiple patient cases. We have defined requirements for the enhancement of the MDHT utility, but their implementation has not been completed.

Our process for assessing clarity relies on the transformation of the CDA xml document into a legible web page. This is performed by a transform – an XSLT template – that renders the xml as html for a web browser to display. The selection of this transform is significant. The CDA Release2 specification includes a simple transform, based on rendering the narrative blocks of the CDA-based document. The VA has built another type of style sheet that renders the information by parsing the structured entries. This more functional approach addresses several issues of optionality and alternate permissible patterns, as observed among VA exchange partners. Furthermore, as mentioned above, this method allows for aggregated views of data domains across multiple sources of data: e.g., one allergy list that consolidates data from the VA EHR and from external sources. These solutions are locally useful, but the CDA specification stipulates that a document cannot rely on a particular transform for faithful presentation. All enhancements built into the transform will have to be fed back to the standards development process (i.e., at HL7) to ensure a common understanding of these ambiguous points.

Sampling

We sampled documents retrieved by the VA from Exchange partners via the eHealth Exchange program in response to queries made by VA clinicians for the purpose of treatment. We tested from one to three documents from each partner, intending to support ongoing, operational monitoring of the quality of data in the eHealth Exchange. We selected documents that had already been requested and received: we were assessing actual operations, as experienced by clinician users, not requesting de novo data to test functionality. All analysis was conducted by trained analysts on systems behind the VA firewall to protect patient confidentiality.

Because we were interested in how participants populated their documents, we did not select files at random, but rather selected those document instances with the most data. As a result, the assessments do not represent an absolute proportion of well-formed data instances provided by a partner, but the proportion of defined elements for which the partner demonstrates the ability to form properly. If a partner had many documents with few sections populated and a few documents with many sections populated, we assessed the latter.

A similar interest guided our scoring within documents. We did not want a preponderance of data in a field with many instances to drown out the quality measure of a field with few instances.

Error Classification

Data quality errors were detected in three ways. First, required fields are counted as errors if missing: optional and conditional fields are not. Second, some errors can be detected by conformance tools, e.g., terminology conformance. And third, human review will identify cases where the document may conform to the technical specification but still fails to provide clear meaning to a reader.

We also endeavored to categorize errors by cause. The general assumption is that an error is the responsibility of the sender, and this is true for all conformance errors. However, there are three possible sources of clarity issues: 1) the author or sender, 2) the stylesheet used to present the data to the reviewer, and 3) in some cases the specification is not sufficiently detailed to be unambiguous, and consequently partners may make different assumptions, each of which conforms to the specification. For any clarity issue, we had to inspect the underlying xml code to determine which of these the case was.

If the issue seemed to be ambiguity in the specification, we identify the ambiguity so that stakeholders can propose new rules — whether to refine the standard specification or to share among eHealth Exchange participants who can then implement them in their own data quality monitoring.

Source errors are reported to the document sender. Transform errors are reported to the VA EHR GUI (VistaWeb) development team so that the assumption can be corrected. Ambiguities in the specification (and recommendations
for their resolution, if appropriate) are reported to the HL7 Structured Documents working group and to the Healtheway Specifications Factory workgroup.

**Scoring**

The partner richness score is a percentage defined as the average number of sections provided in the sample set of selected documents divided by the total number of sections defined by the specification.

The document quality score is the average of quality scores for all of the elements in the document. In a given document instance, each data element gets a score \( S_e \) consisting of the total error-free instances divided by the total expected instances.

\[
S_e = \frac{\text{count of present & correct instances}}{\text{count of expected instances}}
\]

For singular elements, this value is either one or zero, but for repeatable elements it is calculated. A document listing three allergies, but only two with reaction types (a required field for a C32 allergy), would receive a score of 0.67 for the reaction type data element. This way, we can capture effects of intermittent errors without ignoring evidence of appropriate integration.

The document quality score \( S_d \) is the average of these data element scores.

\[
S_d = \text{Average} (S_e)
\]

This average weights all elements equally, so that a heavily populated section does not have undue influence on our understanding of the sender’s capabilities.

The construction of the score in this compositional manner allows scores to be calculated not only for individual documents, but for any slice of information—partners, partner classifications (by size, geography, onboarding stage, etc.), data elements, sections, time periods, etc. It can be used not only for static scores, but also for trends, outliers, fraud detection, or any other perspective that might assist with the continuous improvement process.

The semantic interoperability score is the proportion of coded elements correctly formed, by element, not instance. This distinction is important because a document may contain a code that is incidentally correct. One document, for instance, uses “mg” for milligram, which appears to be the correct Unified Code of Unit of Measure (UCUM) code, but it also uses “K” for thousand, whereas the UCUM meaning of “K” is degrees Kelvin. This document is using local unit codes, some of which happen to be identical to the standard codes. If any errors are detected, all codes for the field must be considered suspect.

Because the tooling for evaluation is still under development, we only have preliminary quantitative assessment of the semantic interoperability score, based on manual counts. All counts are noted in a specially created Excel spreadsheet.

**Results**

![Data Quality and Interoperability Levels](image)

- **Data Richness (-8%)**
- **Semantic Interoperability (+3%)**
- **Data Quality (+3%)**
Figure 2: Example CDA Document Assessment Scores

Figure 2 shows an example partner assessment. The assessed partner’s documents included slightly under half the number of sections specified by the C32 (data richness = 45%). While overall quality scored 70%, coded elements as a group fared significantly worse (26%). Parenthetical figures indicate change from the partner’s previous scores.

Overall VLER partners richness scores ranged from 12% to 82%, with an average of 55% and median of 59%. When richness is low, it is unclear whether the sender is incapable of sending the missing sections or the data is simply not applicable to the patient in question. Automated operational inspection of greater volumes of documents should provide benchmarks for this judgment.

![Count of Richness Scores](image)

Figure 3: Histogram of CDA documents richness score across VA Exchange partners

The manual quality assessment process requires appreciable amounts of time. Due to the small sample size, we are able to provide only qualitative assessments of document quality. In addition, we have been unable to demonstrate a clear trend of quality improvement in incoming documents, given the short time range. We do know that individual sending institutions have responded to our feedback by addressing issues, but this knowledge has not been generalized to the community at large, so any new partner assessed is likely to have quality scores similar to the population average.

We identified several common conformance issues. The existence of these issues suggests that the onboarding process is not sufficient to ensure ongoing operational data quality, and that operational monitoring should be a critical part of the national interoperability infrastructure.

The issues most commonly identified include the following:

1. **Missing required element** (blank field), as required by the C32 specification. Common areas include laboratory result interpretations & reference ranges, encounter comments, allergy reactions and confirmation dates, and medication prescription numbers, dates, and sigs.
2. **Wrong terminology** (populated but with a code not from the specified terminology). Common errors concern units of measure and status values following record status (e.g., “complete”) rather than clinical status (e.g., “resolved”). Local codes and labels for encounters and procedures are often too vague or too organization-specific to be useful, though standard values are not required by the C32.
3. **Misplaced information.** Common errors include stipulation of dose magnitude in medication quantity (which should contain dispensed quantities) and inclusion of immunizations in the medication section. Less
clear are inclusion of radiology results in a results section, and laboratory qualifiers such as patient challenges or blood bank product identifiers.

These errors—clearly stipulated in the specification, yet commonly mis-implemented in incoming documents—confirm that the onboarding process is not sufficient to ensure operational conformance. There is overlap between this list and that provided by D’Amore et al.’ Our “misplaced data” is very similar to their “inappropriate organization,” and our “wrong terminology” is a large subset of their “terminology misuse.” We do not adopt the category “incorrect data,” as we feel the nature of the error can usually be specified. The “Heterogeneity” classifications from that paper are only represented here where an issue creates an error or a clarity issue—e.g., where, in their example, the sporadic reference to narrative text causes a required element to be missing.

Other errors common enough to require their own classification, include

4. **Terminology codes too general to be useful**; e.g., ICD-9 799.9 “Other unknown and unspecified cause of morbidity and mortality.” Usually, this indicates a free-text entry.

5. **Formatting errors**, including the concatenation of type codes (“home,” “work”) or other unidentified text with displayed text.

6. **Duplicated data**—elements that are either duplicated or are valid but so similar to other valid elements that they appear to be duplicates. This occurs more readily when elements are classified at a higher level in a terminology, where a more detailed classification might allow discrimination (see # 4).

We identified a smaller number of issues with the stylesheet used to present documents to clinicians. These findings are fundamentally internal to the VA; we share them in case others have similar issues.

1. Patient names may be repeated or concatenated with aliases.

2. Blood pressure readings are concatenated for a conventional “Systolic/Diastolic” display, but units are omitted.

3. The transform presents scientific notation in an unconventional and potentially ambiguous fashion. Most of the numbers in figure 4 are presented without symbols to orient the reader to the fact that this is scientific notation: “0.4 10” could easily be read as “0.410.” (This presentation issue is in addition to the sender error of not sending the UCUM unit—in the case of monocytes, “%”—making it difficult to compare, e.g., monocyte and granulocyte measurements, as well as the sender error of pairing “count” test names with percentage result values.)

<table>
<thead>
<tr>
<th>%BASO</th>
<th>0.4 %</th>
<th>0.0 - 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>#LYMPH</td>
<td>1.10</td>
<td>Low 1.3 - 2.9</td>
</tr>
<tr>
<td>#MONO</td>
<td>0.410</td>
<td>Normal 0.3 - 0.8</td>
</tr>
<tr>
<td>#GRAN</td>
<td>4.0 10</td>
<td>Normal 2.2 - 4.8</td>
</tr>
<tr>
<td>#EOSIN</td>
<td>0.11</td>
<td>Normal 0.00 - 0.30</td>
</tr>
<tr>
<td>#BASO</td>
<td>0.10</td>
<td>Normal 0.00 - 0.10</td>
</tr>
<tr>
<td>Erythrocytes [#/volume] in Blood by Automated count</td>
<td>4.58 10</td>
<td>Normal 4.00 - 6.00</td>
</tr>
</tbody>
</table>

**Figure 4**: Examples of ambiguous scientific notation in the lab section of a CDA document

The most important findings are those related to specification ambiguity. These issues cannot be solved by testing and feedback loops internal to any organization; they must be solved by consensus within the interoperability community. Given the timeframe necessary to modify a standard, interim measures could be taken within interoperability communities (e.g., Healtheway) by codifying rules that we expect or recommend be adopted by HL7 Structured Documents Workgroup (SD) or C-CDA. This approach has informed the adoption of the “Bridge c32”—a specification used by Healtheway to reduce optionality and ambiguity within its community of participants while
at the same time increasing the richness of the data—as well as the development of the C-CDA. It should be extended to leverage the collective experience of all participants.

These issues include questions about the validity of included data, how to format information consistently, and several questions regarding heterogeneity in the results section.

Questions about the validity of data:

1. What is the temporal scope of a summary document? How far into the past should it reach, and how does this rule vary per clinical topic (e.g., medications, presumably limited, vs. allergies, presumably without limit).
2. May a summary document include data from other sources? From the requestor?
3. If data has been corrected, how is this indicated?
4. For Health Information Exchanges, should the custodian be the HIE or the originating facility or the system that captured the information (e.g., ADT)? A clinician may be more interested in the facility. Documenting data provenance is an area where further specification is needed.
5. How specific should textual information be? Can providers be listed as "historical provider"; or can sourced institutions be listed as abbreviations?
6. Must medications include specific information (e.g., dose), or is the clinical drug alone sufficient?
7. How should fields that are not applicable be represented? For example, in a record of “no known allergies,” how should the allergy reaction be consistently represented? As blanks? As “not applicable”?

Questions about the C32 results section:

1. How should result sections support different kinds of results—hematology and chemistry tests, microbiology tests, radiology reports, etc.?
2. How should ambiguous kinds of acts be listed? Is an x-ray a procedure or a result? What guidelines should be given to implementers regarding what goes into the Results section vs. the Procedures section?
3. How should sparsely populated result qualifiers be represented? E.g., patient challenges, blood bank products, method used, etc.
4. Should reports and clinical notes (e.g., radiology and surgery reports, progress notes, consult notes, discharge summaries, H&Ps) be included in a health summary?
5. If a result itself is an interpretation (e.g., “present”), should there be a redundant value in the interpretation field? An explicit null (e.g., “Not applicable”)? A blank?

Questions about the display text of data:

1. Should a coded element display the preferred term from the code system or a local or manually typed term? Should a code always be represented by the same term in a document? In a facility?
2. When a data element is described with more than one representation (a standard code, a translation code, and an original text), which text should be displayed to a clinician?
3. How should LOINC names be presented? The formal names are often too detailed for clinical use and the short names too abbreviated to be easily understood. Abbreviations affect all sections (e.g., ‘DMII WO CMP NT ST UNCNTR’ for a problem name, ‘O’ for an Encounter type).
4. Is there a problem with appending a code to a coded concept display term (e.g., appending the ICD9 code after the diagnosis name)? Should it be encouraged?
5. Should empty sections be allowed or prohibited? Does an empty section constitute useful information? Should it be displayed as ‘no data provided’, ‘no data available’, ‘no data exist for the date range selected’?
6. How can problem or lab lists be kept to a manageable length? In general, how can the relevant data be made more clear to the receiver?
7. Should the community enforce that the information in the narrative blocks and structured entries be equivalent? If a sender expects the narrative text, as the “attested” data, to be understood, whereas the receiver constructs the human display from the structured data, how can we know that they are semantically
equivalent? Should there be more prescriptive recommendation that the clinician display always be derived from the narrative blocks?

As we consider these questions, we observe that some of them can, once answered by the consensus of stakeholders, be implemented as verifiable rules in the evolving standard. Classification of laboratory information, for instance, should be done once, codified, and implemented. Other questions, however, will require a more nuanced response. Determination of the appropriate temporal or clinical scope of a summary, for instance, will likely be the result of an ongoing discussion among practicing clinicians. The solution to such a question is unlikely to be encoded in a conformance rule, at least in the near future. Instead, practicing clinicians will need a feedback channel to provide insight to document senders. Whether such a loop leverages the expertise of the clinician requesting the document in the first place or that of a dedicated review team remains to be seen, but either way, it should minimize the risk of imposing a burden on clinicians by supporting feedback as terse or detailed as the clinician feels is warranted.

Conclusion

In this paper, we share the VA experience with data quality surveillance of eHealth Exchange participants C32 documents. This surveillance is conducted with real patient data and produces several scores, including data richness, data ‘correctness’, and semantic level.

We offer five recommendations from this experience.

First, we have identified common issues that seem to result from ambiguity in the C32 specification. When the community has moved largely to the C-CDA format, we expect these issues will continue to cause problems, to a lesser degree, but we hope that our experience provides some of the data needed to ameliorate those issues.

Second, we show the importance of continuous surveillance of data quality in a production environment after the certification and onboarding phases are completed. It may be appropriate that certification be a necessary condition for participation in Healtheway, but it is not sufficient to guarantee quality. Each participating organization should consider a C-CDA quality monitor running within their firewall.

Third, we demonstrate that automation can enhance the understanding of data quality, and we support continued investment in tooling for this purpose. While we have seen enough documents to have confidence in our qualitative judgments, we look forward to having the ability to monitor these issues over larger samples, to do so in near-real time, and to use this data in operational feedback to improve the quality of care. These automated tools would come with a set shared public rules that each organization can implement in their environment to continuously assess and alert about the data quality index of the information exchanged. Interoperability is enhanced only when all exchange participants cooperate in monitoring and improvement. One organization cannot do it alone.

Fourth, we have identified several items that may not be amenable to automated testing. Processes for monitoring these issues will require human intervention. Initially, these processes can be conducted based on sampling, but methods should be assessed for capturing operational feedback.

Finally, we recommend that the efforts to identify ambiguities be centralized to support continuous improvement across the Healtheway eHealth Exchange community. As the questions we raise are answered and built into the specification, it will become incrementally more useful, but the list will continue to grow for some time to come, and a governance mechanism will be needed to manage their resolution in ways that continue to support interoperability. This mechanism should support reporting of issues, consensus on solutions, methods for sharing examples, and channels for advocating solutions to standards development groups.

References

