Reconciling Disparate Information in Continuity of Care Documents: Piloting a System to Consolidate Structured Clinical Documents

Authors

1. Masoud Hosseini, MSc, Ph.D. - Corresponding Author

   Department of BioHealth Informatics, School of Informatics and Computing at Indiana University-Purdue University Indianapolis

   Walker Plaza (WK)

   719 Indiana Avenue, WK 117

   Indianapolis, IN 46202, USA

   Email: massood.hosseini@gmail.com

2. Josette Jones, Ph.D.

   Director, Department of BioHealth Informatics, School of Informatics and Computing, Indiana University-Purdue University Indianapolis

3. Anthony Faiola, Ph.D.

   Professor and Head, Biomedical and Health Information Sciences, College of Applied Health Sciences, the University of Illinois at Chicago

4. Daniel J. Vreeman, PT, DPT, MSC

   Associate Research Professor in Medicine, School of Medicine, Indiana University

5. Huanmei Wu, Ph.D.

   Chair, Department of BioHealth Informatics, School of Informatics and Computing, Indiana University-Purdue University Indianapolis

6. Brian E. Dixon, MPA, Ph.D, FHIMSS

   Associate Professor, Department of Epidemiology, Richard M. Fairbanks School of Public health, Indiana University-Purdue University Indianapolis

---

This is the author's manuscript of the article published in final edited form as:

Hosseini, M., Jones, J., Faiola, A., Vreeman, D. J., Wu, H., & Dixon, B. E. (2017). Reconciling Disparate Information in Continuity of Care Documents: Piloting a System to Consolidate Structured Clinical Documents. Journal of Biomedical Informatics, 74, 123-129. https://doi.org/10.1016/j.jbi.2017.09.001

Abstract

Background: Due to the nature of information generation in health care, clinical documents contain duplicate and sometimes conflicting information. Recent implementation of health information exchange (HIE) mechanisms in which clinical summary documents are exchanged among disparate health care organizations can proliferate duplicate and conflicting information.

Materials and Methods: To reduce information overload, a system to automatically consolidate information across multiple clinical summary documents was developed for an HIE network. The system receives any number of continuity of care documents (CCDs) and outputs a single, consolidated record. To test the system, a randomly sampled corpus of 522 CCDs representing 50 unique patients was extracted from a large HIE network. The automated methods were compared to manual consolidation of information for three key sections of the CCD: problems, allergies, and medications.

Results: Manual consolidation of 11,631 entries was completed in approximately 150 hours. The same data were automatically consolidated in 3.3 minutes. The system successfully consolidated 99.1% of problems, 87.0% of allergies, and 91.7% of medications. Almost all of the inaccuracies were caused by issues involving the use of standardized terminologies within the documents to represent individual information entries.

Conclusion: This study represents a novel, tested tool for de-duplication and consolidation of CDA documents, which is a major step toward improving information access and the interoperability among information systems. While more work is necessary, automated systems like the one evaluated in this study will be necessary to meet the informatics needs of providers and health systems in the future.

Keywords

De-duplication, Consolidation, Clinical Document Architecture (CDA), Continuity of Care Document (CCD), Health Information Exchange (HIE), Health Level Seven (HL7)

BACKGROUND AND SIGNIFICANCE

Many factors diffuse health information among multiple, heterogeneous information systems, including variations in insurance coverage, visits to multiple providers, and increasing use of specialty care (1). Information fragmentation makes clinical workflow inefficient, creates barriers to care coordination, and leads to over-utilization of healthcare through duplicative testing or imaging (2, 3).

Health Information Exchange (HIE) can connect fragmented, disparate health information systems and provide timely, relevant information to providers (4-6). The Medicare and Medicaid Electronic Health Record (EHR) Incentive Program regulations, known as "meaningful use," from the U.S. Centers for Medicare and Medicaid Services (CMS) underscore the importance and encourage the use of HIE to support better care (7). Exchanging key clinical data, transmitting a summary care record, e-prescribing, integrating laboratory results, and immunization reporting are all different forms of HIE required to demonstrate the meaningful use of EHR systems (8).

Health Level Seven (HL7) is an international standards development organization that creates standards for exchanging clinical and administrative data among healthcare information systems (9-11). The CMS EHR Incentive Program rules adopted the HL7 Consolidated CDA (C-CDA) Implementation Guide as "the sole standard for exchanging summary care records" (12). The C-CDA standard (13) is not a singular type of document or message but rather a framework of templates for multiple clinical documents, including the Continuity of Care Document or CCD, Clinical Oncology Treatment Plan and Summary, and Emergency Medical Services Patient Care Report. The C-CDA framework as a method to represent information is comparable to the concept of archetypes for representing data in EHR systems (14-16). Of the available CDA templates, the CCD is one of the more popular documents generated and exchanged by healthcare facilities (17-19). Its popularity stems, in part, from its early adoption by the eHealth Exchange (formerly known as the Nationwide Health Information Exchange) as the primary document for HIE across routine clinical use cases such as medication reconciliation (20) and care coordination (21).

Although information sharing using CCDs is an improvement over fragmented data silos, the rapid and repeated exchange of CCDs for patients has created new challenges. Private as well as community-based HIE networks (22) now exchange multiple CCDs for the same patient from different sources, yet they may not decompose the discrete data in each document into a unified clinical summary. For example, within the U.S. Veterans Administration (VA), which is exchanging CCDs with hundreds of non-VA providers via dozens of community-based HIE networks (23), clinicians are provided access to a list of CCDs for a patient gathered from several non-VA sources. This requires VA clinicians to sift through multiple documents that include potentially duplicative or conflicting information. Similar reports and

experiences have been reported in other HIE networks (24, 25). Without an efficient and effective method for de-duplicating redundant patient data, providers everywhere will be forced to review lengthy and redundant information spread across multiple documents. This will create further information overload and usability challenges for EHR systems, which are already criticized for not supporting clinical workflow (26, 27).

To combat information overload, we developed a system to consolidate and de-duplicate multiple CCDs for a single patient. The system was designed for use by an HIE network, and we previously described its design and a preliminary evaluation using simulated data (28). Encouraged by the preliminary results, we continued to evolve the system. In this article, we describe improvements made to the system and present the results of an evaluation using real-world CCDs sampled from a large, community-based HIE network.

OBJECTIVES

The purpose of this study is to evaluate the accuracy of a system designed to consolidate or de-duplicate disparate information contained in multiple CCDs for the same patient exchanged within a production HIE network. We further seek to identify the main causes of inaccurate consolidation and investigate the differences in data representation across a set of real-world CCDs.

MATERIALS AND METHODS

To evaluate the system that consolidates information across multiple CCDs for the same patient, we compared the output of information reconciliation by the system with that of manual reconciliation for a sample of real-world CCDs while also measuring system performance. For instances where the system was unable to reconcile two data entries, we classified the errors to identify areas for future system development and research. Our study was approved by the Institutional Review Board at Indiana University (1503069512) and conducted as a component of the primary author's doctoral dissertation research.

System Description

The system consists of three major components: 1) a Consolidation Engine that receives multiple CDA documents for a single patient through an application programming interface (API) and executes a series of rules to de-duplicate data across input documents then merge the information into a single, superset CDA document for the patient; 2) an Audit component that identifies and logs system events; and 3) a configuration component that enables implementers to configure the rules executed by the system. The system uses a RESTful API and web application to enable users to configure the desired rule sets they wish to enable during the consolidation process. Multiple CDA documents are input into the

Consolidation Engine, and the system outputs a de-duplicated, merged CDA document after executing the configured rules enabled by the user or calling application. The output represents the union of the input documents. During execution, audit information is captured into a NoSQL database. Figure 1 abstractly depicts the system. A more complete description of the system is provided elsewhere (28).

Figure 1. Abstract depiction of the system. Given an input of multiple CDA documents (D_n) for the same patient, the system finds all unique data entries. These entries, which represent the union of the data across the set of documents, are output as a single, consolidated CDA document.

The system de-duplicates information by comparing data entries (e.g., code, code system, healthcare provider name) across the same section (e.g., Problems, Medications, Allergies) of the input CDA documents. Semantically equivalent entries (e.g., exact text match, same ICD code) are reduced to a single entry in the appropriate section of the output CDA document. Unique entries are retained in the output CDA document. The output of the system represents a single, consolidated superset of the input CDA documents.

Since originally describing the system, the rules engine was enhanced to more closely examine the data elements in CDA documents. For example, “nullflavor” and “negationInd” are attributes that can alter the meaning of a data element, so we enhanced several rules to de-duplicate data while considering these attributes. In addition to the semantic standards recommended for use in CDA documents (e.g., ICD, SNOMED, RxNorm), translation codes may be used to indicate multiple, alternative semantic representations for a single entry. The system now uses translation codes, when present, to identify semantic equivalence by comparing entries to available codes included in the set of input CDA documents.

Study Design

To evaluate the system, we used a corpus of CCDs exchanged by health systems participating in a large HIE network. First, we consolidated the corpus and de-duplicated the information manually. Next, we ran the corpus through the system for automated consolidation and de-duplication of information.

We then compared the results between the manual and automated approaches. During this comparison, we classified errors (missed opportunities for de-duplication) to better understand how we might refine the system going forward.

Methods for Data Acquisition

Our corpus of real-world CCDs were sampled from the Indiana Network for Patient Care (INPC), one of the nation's largest and longest tenured HIE networks. The INPC includes more than 60 hospitals representing multiple health systems as well as community hospitals, large physician practices, independent radiology centers, and other health care organizations. The INPC is managed by the Indiana Health Information Exchange (IHIE), which serves over 25,000 physicians, and its data repository includes over 10 million patients (29, 30).

Staff at IHIE provided 176,169 CCDs the network received between 07/12/2014 to 11/06/2015 (484 days) as a component of Stage 2 meaningful use compliance. In other words, these were CCDs sent from a hospital or outpatient facilities following a transition of care and which represented a single encounter. We then grouped the CCDs into folders based on patient identifiers provided in each CCD. This allowed identification of unique patients with multiple CCDs for sampling; it did not make sense to include patients with only one CCD.

After grouping CCDs, we identified a total of 30,370 unique patients for which there were at least two CCDs; overall there were 86,184 unique CCDs. Before sampling CCDs, we categorized patients into classes based on the number of CCDs (2 to 36) in the folder. This enabled sampling from classes with proportionately more unique patients to balance selection. From each class, we randomly sampled patients to construct a corpus representing 50 total unique patients. The final corpus included 522 CCDs for 50 unique patients with a range of 2 to 36 CCDs per patient. The CCDs per patient range from 2 to 36 with mean of 2.84 and standard deviation of 10.9.

Manual Consolidation and De-Duplication

The entire corpus of 522 CCDs was manually reviewed by the author MH, entry by entry, for three key sections of the CDA standard: Problems, Allergies, and Medications. These sections were targeted because they contain key information necessary to perform important clinical tasks, such as medication reconciliation, drug-drug interaction checking, and drug-allergy checking. As each section was reviewed,

data duplication across the set of CCDs for a given patient was recorded for analysis. A manually constructed, consolidated CCD along with metadata (e.g., number of CCDs, number of entries, time to review) was saved to the patient's folder for analysis.

Automated Consolidation and De-Duplication

Using a test harness, or simple user interface designed to enable testing of the system, each batch of CCDs for the 50 unique patients was input into the system for automated consolidation and data de-duplication. The system outputs a final, consolidated CCD along with runtime statistics. The same metadata (e.g., number of CCDs, number of entries, time to process the batch) were saved to the patient's folder along with the consolidated CCD for analysis.

Methods for Data Analysis

To measure the accuracy of the system, we compared the results of the manual consolidation to that of the automated consolidation using the system. We considered the manual consolidation process as the 'gold standard,' because manual review is how clinicians must today reconcile information across CCDs when reviewing a list of patient documents. By comparing the manual results to those of the system, we measured how well the system could correctly identify that two or more data entries were semantically equivalent across the batch of input CCDs and could therefore be reduced to a single entry in the final, consolidated CCD. For instance, if there were seven duplications identified during the manual review, and 5 detected by the system during the automated process, we would conclude that the accuracy of the system is 71%. The accuracy of the manual consolidation is 100% as it is considered as the 'gold standard'.

The comparison between manual and automated consolidation was performed for each unique patient. We further recorded the number of entries across each batch of CCDs before and after consolidation, including the number of entries per each of the three sections targeted by the system.

In addition to counting entries and calculating accuracy rates, MH further examined and classified each mismatch between the manual and automated consolidation. For example, if the number of final entries in the allergy section was different between the final, consolidated CCDs for the manual and automated runs, then MH examined and classified which entries were not consolidated. The errors and their classifications were recorded for summary and presentation.

Finally, we examined the overall performance of the system. System performance was evaluated based on the total time necessary to process the 522 input CCDs. We examined the average time to process a CCD as well as the minimum, maximum, and average file sizes.

Results

Table 1 summarizes the result of consolidation on the 522 CCDs. We observed significant duplication of information across the sample of CCDs as both manual and automated consolidation reduced the number of individual entries by over 9000 (80%). The medication section contained the highest number of unique entries while the problem and allergy sections repeated items more frequently.

With respect to accuracy, the system agreed with manual consolidation by 95% overall, and agreement was at least 80% within each section of the CCD. Accuracy was strongest in the problem section (99%) followed by the medication section (92%). Accuracy was weakest in the allergy section (87%).

	Problem Section	Allergy Section	Medication Section	Overall
Manual Consolidation				
No. Entries in Input CCDs (N=522)	6,531	1,845	3,255	11,631
No. Entries in Consolidated CCDs (N=50)	659	150	1,390	2,199
Reduction (%)	89.9%	91.9%	57.3%	81.1%
No. Duplications	5,872	1,695	1,865	9,432
Automated Consolidation				
No. Entries in Input CCDs (N=522)	6,531	1,845	3,255	11,631
No. Entries in Consolidated CCDs (N=50)	711	370	1,545	2,626
Reduction (%)	89.1%	79.9%	52.5%	77.4%
No. Duplications	5,820	1,475	1,710	9,005
Accuracy (# Automated Duplicates / # Manual Duplicates)	99.1%	87%	91.7%	95.5%

Table 1. Total number of entries in each section of all CCDs.

Manual consolidation of 11,631 entries required approximately 150 hours of effort, whereas automated consolidation occurred in 3.3 minutes (average 0.38 seconds for each CCD). The size on disc for the input

of 522 CCDs was 265 MB. After consolidation, the output 50 CCDs required 26.2 MB of disc storage, representing a 90% decrease.

Inaccuracy in system deduplication

Despite high rates of consolidation and de-duplication, the system did not perfectly reconcile all duplicate entries. Automated consolidation detected less duplication than manual methods. We did not observe any false positives where two entries were classified as duplicates by the system but were found in manual review to be unique. This was expected, because the system is designed to prioritize specificity.

Two types of false negatives were observed where the system did not classify two entries as duplicates but human review found them to be duplicates. The first group of false negatives occurred when the system was not able to compare data elements because the values were free text instead of coded entries. All of the mismatches in the problem and allergy sections were due to the use of the free-text section of the CCD rather than a semantic terminology code. The system bases its comparison of entries based on structured, encoded concepts, and it relies on standards such as SNOMED, RxNorm, or ICD. Yet sometimes coded entries are missing as in Figure 2 where an allergy entry does not include any code. Instead there is simply a reference to the free-text section. Figure 3 on the other hand, presents another allergy entry that is accurately de-duplicated by the system, because it includes two different codes for "Codeine" to which a patient is allergic.

```
<observation classCode="OBS" moodCode="EVN">
  <templateId root="2.16.840.1.113883.10.20.22.4.7" />
  <id root="                                        " />
  <code code="ASSERTION" codeSystem="2.16.840.1.113883.5.4" codeSystemName="HL7 ActCode" displayName="Assertion" />
  <statusCode code="completed" />
  <effectiveTime>...</effectiveTime>
  <value xsi:type="CD" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED CT" code="419199007" displayName="Allergy to substance" />
  <author>...</author>
  <participant typeCode="CSM" contextControlCode="OP">
    <participantRole classCode="MANU">
      <playingEntity classCode="MMAT">
        <code>
          <originalText>
            <reference value="#ALLERGEN       " />
          </originalText>
        </code>
      </playingEntity>
    </participantRole>
  </participant>
  <entryRelationship typeCode="SUBJ" inversionInd="true">...</entryRelationship>
  <entryRelationship typeCode="MFST" inversionInd="true">...</entryRelationship>
</observation>
```

Figure 2. An allergy entry without code for substance to which patient is allergic to.

```
<observation classCode="OBS" moodCode="EVN">
  <templateId root="                                    " />
  <id root="                                          " />
  <code code="ASSERTION" codeSystem="2.16.840.1.113883.5.4" codeSystemName="HL7 ActCode" displayName="Assertion" />
  <statusCode code="completed" />
  <effectiveTime>...</effectiveTime>
  <value xsi:type="CD" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED CT" code="416098002" displayName="Drug allergy" />
  <author>...</author>
  <participant typeCode="CSM" contextControlCode="OP">
    <participantRole classCode="MANU">
      <playingEntity classCode="MMAT">
        <code code="2670" codeSystem="2.16.840.1.113883.6.88" codeSystemName="RxNorm" displayName="Codeine">
          <originalText>
            <reference value="#ALLERGEN       " />
          </originalText>
          <translation code="d00012" codeSystem="2.16.840.1.113883.6.314" codeSystemName="multum-drug-id" displayName="codeine" />
        </code>
      </playingEntity>
    </participantRole>
  </participant>
  <entryRelationship typeCode="SUBJ" inversionInd="true">...</entryRelationship>
  <entryRelationship typeCode="MFST" inversionInd="true">...</entryRelationship>
</observation>
```

Figure 3. An allergy entry with two codes for Codeine to which patient is allergic.

The second type of false negative occurs when an alternative coding scheme is used in lieu of a primary coding standard. This type of false negative was observed in the medication section of the CCDs and accounted for all of the mismatches between the manual and automated consolidations. Although the meaningful use regulations stipulate that RxNorm should be used to identify medication entries, the HL7 CCDA standard allows for a nullFlavor value of "OTH" to represent that, instead of a primary coding scheme, an alternative (or other) coding scheme is used to identify the medication entry. We observed several instances where the OTH value was used and a translation code (e.g., Multum Drug ID) was provided as the only way to identify the medication entry. An example is shown in Figure 4. In these instances, the system keeps both entries because the alternative code is not necessarily unique and therefore unreliable. For instance, the mediation "24 HR diltiazem Hydrochloride Extended Release Capsule" has two dosages of 240 MG and 180 MG. The Multum Drug ID for both dosages is the same (D00045), while the RxNorm codes are distinct (830837 and 830845). Although these medications can be correctly de-duplicated using their RxNorm codes, de-duplicating using only Multum Drug IDs might be difficult and lead to false positives. Therefore the system does not de-duplicate using translation codes for OTH entries where there is no primary coding scheme.

```
<code nullFlavor="OTH">
  <originalText>
    <reference value="#MEDPROD6400065107" />
  </originalText>
  <translation code="d03768" codeSystem="2.16.840.1.113883.6.314" codeSystemName="multum-drug-id" />
</code>
```

Figure 4. A medication without RxNorm code.

DISCUSSION

Using a random sample of real-world CDA documents exchanged within a large HIE network, we assessed the accuracy and performance of a web service designed to consolidate and de-duplicate CDA documents generated by EHR systems for the same patient. When compared to manual consolidation (gold standard), the automated system achieved reasonable rates of accuracy. Furthermore, system performance was reasonable with a 0.38 seconds per document processing rate. These results provide greater evidence of feasibility beyond our earlier pilot involving synthetic data (31), paving the way for the system to be further refined and implemented within an HIE network.

As policies such as meaningful use continue to push health information systems to generate and exchange CDA-based documents, the need for consolidation tools like the one tested will increase. CDA documents represent snapshots in time of care delivered to a patient, and as patients traverse multiple providers for acute and chronic care the number of CDA-based documents containing fragments of their complete, longitudinal medical record will grow. Both HIE networks and EHR systems need efficient methods for consolidating information to support use by providers who utilize various clinical workflows or to capture "new" or "updated" information into a repository for storage. The high levels of duplication observed across a relatively small subset of CDA documents sent from enterprise EHR systems, especially in the problem list and allergy sections of the CCD, underscores this need. Therefore the findings presented here are useful to both those who may wish to develop or implement similar tools in their health systems or HIE networks.

Building upon our prior work, this study identified three key lessons for consolidating information across disparate CDA-based documents. First, adherence to technical standards is critical to success. Second, mismatches between the coded and free-text sections of a CDA document present unique human factors challenges for consolidation engines. Finally, consolidation engines can identify conflicting information in CDA documents but resolving conflicts remain a challenge.

Adherence to standards critical to success

When health information systems adhere to syntactic and semantic standards, computers can efficiently reconcile data from disparate messages or documents. In this study, we observed that nearly all of the issues encountered by the system were due to missing semantic coding within a structured data element. When a data entry lacked semantic encoding, it was challenging to compare that element to the others included in the set of CCDs under examination. In these cases, the system attempted to leverage local identifiers as well as perform string comparisons on concept names. Yet these methods are imprecise, requiring perfect alignment of text entries. Therefore the system had no alternative but to output duplicates in the final, consolidated CCD.

One approach for the future might be to explore the use of fuzzy matching algorithms and word stemming, techniques used in natural language processing (32-35). While such approaches should be explored, alternatives to semantic encoding are unlikely to be as accurate given the robust and mature nature of available standards (36).

While inaccuracies in de-duplicating concepts were caused primarily by a lack of semantic interoperability, overall we observed strong adherence to both syntactic and semantic standards. None of the CCDs were rejected by the system for failing to conform to the HL7 CDA specifications. This is notable, because there has been suspicion within the health IT industry that real-world messages emanating from commercial EHR systems may not conform as much as expected given 'meaningful use' certification criteria. Furthermore, the majority of structured data elements within the CCDs contained appropriately encoded concepts using terminology standards such as SNOMED, LOINC and RxNorm. These are the terminology standards required for 'meaningful use' certification (37), and it was pleasant to find their use among CDA-based documents generated by real-world, enterprise EHR systems. A prior analysis by Dixon et al. (38) of real-world messages in the INPC demonstrated low adoption of semantic standards. Thus when standards were used the system encountered few problems parsing the data and consolidating them across multiple documents.

Semantic mismatches present human factors challenges

A smaller cause of inaccuracy but worth noting is the case where two concepts were identified as unique based on their semantic encoding but for which the unstructured, "human readable" sections of the CCD contains the exact same value. In these cases, the consolidation engine appropriately kept both entries for the final CCD. While correct in its determination, this behavior could cause a human factors challenge for providers who view the output CCD (39). When the two unstructured entries are rendered in a user interface, the provider may interpret them as being duplicate information. Noticing duplication would be a minor inconvenience in what is otherwise a significantly scaled down, consolidated document. Therefore, this type of error may be useful for consolidation engines to flag in order to raise awareness with data producers. Ideally data sources will begin to more appropriately distinguish entries using both the structured and unstructured sections of CDA documents.

Conflicting data present consolidation challenges

A final challenge to note is conflicting information across multiple CDA documents. Currently the system can detect conflicts across sections such as the problem or medication list. For example, one CCD may show that a patient is allergic to penicillin but the other CCD reports the patient has no known allergies. In this case the system generates the consolidated CCD with information that states patient is allergic to

penicillin. However, the system is challenged to appropriately resolve such conflicts. The current rule sets are configured in these instances to allow, for example, an entry for the presence of an allergy and the notice of "no known allergies" as two distinct entries in the final allergy section. More sophisticated approaches are necessary to develop for resolving complex conflicts in the data observed across multiple CCDs. This is an area for future research by our team as well as others developing tools for consolidating and presenting data to providers drawn from multiple sources.

Limitations and Future Development

This study possesses three limitations that suggest future development and work on the consolidation of information in clinical documents. These limitations pertain to the methods used for comparison, semantic overlap across terminology standards, and overall scope of the consolidation system.

This study treated the manual consolidation performed by the lead author (MH) as the gold standard. Humans, however, are not perfect and therefore using only one manual reviewer may have resulted in our methods missing some true or false matches. A single reviewer was used in this study due to limitations of time and budget; the work was performed as part of the lead author's dissertation and did not have external funding. Future testing and refinement of the system will be necessary to improve its performance as well as accuracy.

Although the system utilizes translation codes (along with the canonical standard codes to identify duplications), the de-duplication engine is not able to detect semantic overlaps when different semantic terminologies are present. For example, Lyme disease can be expressed as both a SNOMED CT concept (23502006) and as an ICD-10 CM concept (A.69.2). Approaches that leverage the UMLS Metathesuarus could be explored in future work to take advantage of already cross-linked ontologies.

The Meaningful Use rules refer to the "adoption of solely the Consolidated CDA (C-CDA) standard for summary care records." Yet the C-CDA implementation guide (13) includes 9 different clinical documents of which the CCD is only one. Moreover, this study focused on just three sections of a CCD (problems, medications, and allergies). Because the system is designed in a way to be extensible, future work remains to address all of the C-CDA variants as well as additional sections among the 17 possible sections including laboratory tests and procedures. There are also opportunities to enhance the system to consolidate HL7 FHIR (Fast Healthcare Interoperability Resources) based documents, an emerging standard that some suggest may one day supplant current CDA-based exchange of information.

Conclusion

The use and exchange of CDA-based documents are expected to increase as a larger number of health systems expand their HIE capacity in compliance with HIE policies. As such, providers will gain access to an increasing volume of patient medical records from multiple care settings and health systems. Duplication and conflict of patient information across medical records is therefore inevitable. The potential of having a document consolidation engine as a HIE service will provide clinicians information that is (1) easier to use and understand and (2) more searchable/findable due to its integration of homogenous information into one section. This study represents a novel, tested tool for de-duplication and consolidation of CDA documents, which is a major step toward improving information access and the interoperability among information systems. While more work is necessary, automated systems like the one evaluated in this study will be necessary to meet the informatics needs of providers and health systems in the future.

Funding Statement

This research is not supported by any funding agency, commercial or not-for-profit.

Competing Interests Statement

The authors have no competing interests to declare.

References

1 Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. Journal of the American Medical Informatics Association. 2010;**17**(3):288-94.
2 Schoen C, Osborn R, Squires D, et al. A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health affairs (Project Hope). 2012 Dec;**31**(12):2805-16.
3 Frisse ME, Johnson KB, Nian H, et al. The financial impact of health information exchange on emergency department care. J Am Med Inform Assoc. 2012 May-Jun;**19**(3):328-33.
4 Brailer DJ. Interoperability: the key to the future health care system. Health Affairs-Millwood Va Then Bethesda MA. 2005;**24**:W5.
5 Acker B, Birnbaum C, Braden J, et al. HIM principles in health information exchange. Journal of AHIMA/American Health Information Management Association. 2007;**78**:69-74.
6 Hripcsak G, Kaushal R, Johnson KB, et al. The United Hospital Fund meeting on evaluating health information exchange. Journal of biomedical informatics. 2007;**40**(6):S3-S10.
7 Adler-Milstein J, Bates DW, Jha AK. A survey of health information exchange organizations in the United States: implications for meaningful use. Ann Intern Med. 2011 May 17;**154**(10):666-71.
8 Kuperman GJ. Health-information exchange: why are we doing it, and what are we doing? Journal of the American Medical Informatics Association. 2011;**18**(5):678-82.
9 Hosseini M, Ahmadi M, Dixon BE. A Service Oriented Architecture Approach to Achieve Interoperability between Immunization Information Systems in Iran. AMIA Annu Symp Proc. 2014;**2014:1797-805**.
10 Office of the National Coordinator (ONC). 2014 Edition Release 2 Electronic Health Record (EHR) Certification Criteria and the ONC HIT Certification Program; Regulatory Flexibilities, Improvements, and Enhanced Health Information Exchange. The U.S. Department of Health and Human Services; 2014.
11 Health Level Seven web site. "About HL7". [cited; Available from: http://www.hl7.org/about/index.cfm
12 Medicare and Medicaid programs; electronic health record incentive program--stage 2. Final rule. Federal register. 2012 Sep 4;**77**(171):53967-4162.
13 HL7 Health Level Seven. HL7 Implementation Guide for CDA® Release 2: Consolidated CDA Templates for Clinical Notes. 2012 [cited; Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=379
14 Gonzalez-Ferrer A, Peleg M, Marcos M, Maldonado JA. Analysis of the process of representing clinical statements for decision-support applications: a comparison of openEHR archetypes and HL7 virtual medical record. Journal of medical systems. 2016 Jul;**40**(7):163.
15 Bointner K, Duftschmid G. HL7 template model and EN/ISO 13606 archetype object model - a comparison. Studies in health technology and informatics. 2009;**150**:249.
16 Xu C, Berry D, Stephens G. Using a generalised identity reference model with archetypes to support interoperability of demographics information in electronic health record systems. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2015;**2015**:6834-9.
17 D'Amore JD, Sittig DF, Ness RB. How the continuity of care document can advance medical research and public health. American journal of public health. 2012 May;**102**(5):e1-4.
18 Simonaitis L, Belsito A, Cravens G, Shen C, Overhage JM. Continuity of Care Document (CCD) Enables Delivery of Medication Histories to the Primary Care Clinician. AMIA Annu Symp Proc. 2010;**2010**:747-51.
19 Dolin RH, Giannone G, Schadow G. Enabling joint commission medication reconciliation objectives with the HL7 / ASTM Continuity of Care Document standard. AMIA Annu Symp Proc. 2007:186-90.
20 Simonaitis L, Dixon BE, Belsito A, Miller T, Overhage JM. Building a production-ready infrastructure to enhance medication management: early lessons from the nationwide health information network. AMIA Annu Symp Proc. 2009;**2009**:609-13.
21 Kuperman GJ, Blair JS, Franck RA, Devaraj S, Low AF. Developing data content specifications for the nationwide health information network trial implementations. J Am Med Inform Assoc. 2010 Jan-Feb;**17**(1):6-12.
22 Dixon BE. What is Health Information Exchange? In: Dixon BE, editor. Health Information Exchange: Navigating and Managing a Network of Health Information Systems. Waltham, MA: Academic Press; 2016. p. 3-20.
23 Byrne CM, Mercincavage LM, Bouhaddou O, et al. The Department of Veterans Affairs' (VA)

implementation of the Virtual Lifetime Electronic Record (VLER): findings and lessons learned from Health Information Exchange at 12 sites. Int J Med Inform. 2014 Aug;**83**(8):537-47.
24 Gadd CS, Ho YX, Cala CM, et al. User perspectives on the usability of a regional health information exchange. J Am Med Inform Assoc. 2011 Sep-Oct;**18**(5):711-6.
25 Reis J, MacKenzie L, Soelberg T, Smith J. Assessment of the usability and impact of the Idaho Health Data Exchange (IHDE). Journal of medical systems. 2016;**40**(4):1-7.
26 McDonald CJ, Callaghan FM, Weissman A, Goodwin RM, Mundkur M, Kuhn T. Use of internist's free time by ambulatory care Electronic Medical Record systems. JAMA Intern Med. 2014 Nov;**174**(11):1860-3.
27 Ratwani RM, Fairbanks RJ, Hettinger AZ, Benda NC. Electronic health record usability: analysis of the user-centered design processes of eleven electronic health record vendors. J Am Med Inform Assoc. 2015 Nov;**22**(6):1179-82.
28 Hosseini M, Meade J, Schnitzius J, Dixon BE. Consolidating CCDs from multiple data sources: A modular approach. Journal of the American Medical Informatics Association. 2015;**2016;23(2):317-23**:ocv084.
29 Biondich PG, Grannis SJ. The Indiana network for patient care: an integrated clinical information system informed by over thirty years of experience. J Public Health Manag Pract. 2004 Nov;**Suppl**:S81-6.
30 Overhage JM. The Indiana Health Information Exchange. In: Dixon BE, editor. Health Information Exchange: Navigating and Managing a Network of Health Information Systems. 1 ed. Waltham, MA: Academic Press; 2016. p. 267-79.
31 Hosseini M, Meade J, Schnitzius J, Dixon BE. Consolidating CCDs from multiple data sources: a modular approach. J Am Med Inform Assoc. 2016 Mar;**23**(2):317-23.
32 Chaudhuri S, Ganjam K, Ganti V, Motwani R. Robust and efficient fuzzy match for online data cleaning. Proceedings of the 2003 ACM SIGMOD international conference on Management of data; 2003: ACM; 2003. p. 313-24.
33 Yang Y, Chute CG. An example-based mapping method for text categorization and retrieval. ACM Transactions on Information Systems (TOIS). 1994;**12**(3):252-77.
34 Bhasuran B, Murugesan G, Abdulkadhar S, Natarajan J. Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. Journal of Biomedical Informatics. 2016 12//;**64**:1-9.
35 Robert C, Cousson-Gélie F, Faurous W, Mathey S. Subjective Lexical Characteristics: Comparing Ratings of Members of the Target Population and Doctors for Words Stemming from a Medical Context. Language and Speech. 2016;**59**(4):562-75.
36 Alyea JM, Dixon BE, Bowie J, Kanter AS. Standardizing Health-Care Data Across an Enterprise. In: Dixon BE, editor. Health Information Exchange: Navigating and Managing a Network of Health Information Systems. 1 ed. Waltham, MA: Academic Press; 2016. p. 137-48.
37 2015 Edition Health Information Technology (Health IT) Certification Criteria, 2015 Edition Base Electronic Health Record (EHR) Definition, and ONC Health IT Certification Program Modifications. Final rule. Fed Regist. 2015 Oct 16;**80**(200):62601-759.
38 Dixon BE, Vreeman DJ, Grannis SJ. The long road to semantic interoperability in support of public health: experiences from two states. J Biomed Inform. 2014 Jun;**49**:3-8.
39 Holden RJ, Voida S, Savoy A, Jones JF, Kulanthaivel A. Human Factors Engineering and Human–Computer Interaction: Supporting User Performance and Experience. In: Finnell JT, Dixon BE, editors. Clinical Informatics Study Guide: Text and Review. Zurich: Springer International Publishing; 2016. p. 287-307.

Competing Interests Statement

The authors have no competing interests to declare.

Abstract depiction of the system. Given an input of multiple CDA documents (D_n) for the same patient, the system finds all unique data entries. These entries, which represent the union of the data across the set of documents, are output as a single, consolidated CDA document.

RECONCILING DISPARATE INFORMATION IN CONTINUITY OF CARE DOCUMENTS: PILOTING A SYSTEM TO CONSOLIDATE STRUCTURED CLINICAL DOCUMENTS

Highlights

- A system to automatically consolidate multiple CCD documents for a patient was developed and evaluated.
- Compared to manual consolidation, the system successfully consolidated 99.1% of problems, 87.0% of allergies, and 91.7% of medications.
- This novel tool is a major step toward reducing information overload and improving the interoperability among information systems.