Screening Consolidated Clinical Document Architecture (CCDA) Documents for Sensitive Data Using a Rule-Based Decision Support System

Beatriz H. Rocha[1,2,*]; Deepika Pabbathi[3]; Molly Schaeffer[3]; Howard S. Goldberg[2,3,4]

[1]Wolters Kluwer Health, Waltham, United States; [2]Harvard Medical School, Boston, United States; [3]Partners HealthCare System, Information Systems, Wellesley, United States; [4]Brigham and Women's Hospital, Boston, United States

Keywords

Clinical decision support, meaningful use, health information exchange, continuity of patient care

Summary

Background: The Centers for Medicare & Medicaid Services' Stage 2 final rule requires that eligible hospitals provide a visit summary electronically at transitions of care in order to qualify for "meaningful use" incentive payments. However, Massachusetts state law and Federal law prohibit the transmission of documents containing "sensitive" data unless there is a new patient consent for each transmission.

Objectives: To describe the implementation and evaluation of a rule-based decision support system used to screen transition of care documents for sensitive data.

Methods: We implemented a rule-based document screening system to identify transition of care documents that might contain sensitive data. The transmission of detected documents is withheld until a new patient consent is obtained. The documents that were flagged as containing sensitive data were reviewed in two different time periods to verify that the decision support system was not missing documents or withholding more documents than necessary.

Results: The rule-based screening system has been in regular production use for the past 18 months. During the first evaluation period, 3% of 5,841 documents were identified as containing sensitive data (true-positive rate of 44%). After additional enhancements to the rules, the system was evaluated a second time and 4.5% of 6,935 documents were identified as containing sensitive data (true-positive rate of 98.4%).

Conclusion: The analysis of the system demonstrates that production rules can be used to automatically screen the content of transition of care documents for sensitive data. The utilization of the rule-based decision support system enabled our hospitals to achieve meaningful use and, at the same time, remain compliant with state and federal laws.

Correspondence to:
Beatriz H. Rocha, MD, PhD
Brigham and Women's Hospital
General Medicine
1620 Tremont Street
Boston, MA 02120
USA
Email: bhscrocha@icloud.com
Scopus Author ID: 7005976021

Appl Clin Inform 2017; 8: 137–148
https://doi.org/10.4338/ACI-2016-07-RA-0120
received: July 20, 2016
accepted: December 4, 2016
published: Febrary 8, 2017
Citation: Rocha BH, Pabbathi D, Schaeffer M, Goldberg HS. Screening consolidated clinical document architecture (CCDA) documents for sensitive data using a rule-based decision support system. Appl Clin Inform 2017; 8: 137–148
https://doi.org/10.4338/ACI-2016-07-RA-0120

* The study was conducted while the primary author was a Principal Medical Informatician at Brigham and Women's Hospital

1. Background and Significance

In 2009 the US government issued "The American Recovery and Reinvestment Act" (ARRA) to authorize incentive payments to eligible hospitals to promote the adoption and meaningful use of Certified Electronic Health Record Technology (CEHRT) [1]. The Centers for Medicare & Medicaid Services (CMS) Stage 2 final rule [2] specifies the criteria that eligible hospitals must meet in order to qualify for "meaningful use" incentive payments. One of these criteria specify that eligible hospitals must provide a summary of care record for each transition of care or referral for more than 50% of transitions of care and referrals, and that 10% of such transitions be electronically transmitted. When transmitted electronically, these documents must follow the CEHRT (2014 Edition) requirements, which specify the use of the Health Level Seven (HL7) Consolidated Clinical Document Architecture (CCDA) standard [3].

In order to qualify for these incentives, Partners HealthCare System (PHS) has implemented a computerized process that automatically creates CCDAs for different legacy Electronic Health Record systems (EHRs). These CCDAs are sent to the patient's primary care physician and/or referring provider after the patient is discharged from the hospital. However, Massachusetts state law and Federal law prohibit the transmission of documents that contain "sensitive" data (e.g., human immunodeficiency virus (HIV) results), unless there is an additional specific consent for each transmission (a pre-existing general consent is not sufficient) [4, 5]. In an attempt to prevent the transmission of CCDAs containing sensitive data without specific patient consent, PHS implemented a rule-based document screening system that analyses the content of CCDAs generated during transitions of care and verifies if the documents contain sensitive data or not. CCDA documents identified by the screening system as containing sensitive data are not automatically transmitted. Instead, these documents are marked as containing sensitive data and not sent out electronically. The goal of this process is to achieve balance between protecting patient information and minimizing disruptions during transitions of care [6–9].

PHS is an integrated delivery network with 10 hospitals and several outpatient clinics. Its two founding hospitals, Massachusetts General Hospital and Brigham and Women's Hospital, are Academic Medical Centers affiliated with Harvard Medical School. At the time of this study, the two Academic Medical Centers used distinct "homegrown" (local) EHRs that were certified for the CEHRT 2014 Edition. The other PHS hospitals used different vendor-provided EHRs that were also certified for CEHRT 2014 Edition.

The "CDA Factory" is a homegrown software service integrated with the local EHRs that generates documents compliant with the HL7 Clinical Document Architecture [10] standard. The service is capable of creating several restrictions of the standard, including the "Continuity of Care Document" (CCD) [11], Healthcare Information Technology Standards Panel (HITSP) C32 v2.5 [12], and CCDA [3] documents. The CDA Factory is able to retrieve and automatically translate a wide variety of clinical data types found in the local EHRs. The CDA Factory was used to generate the Academic Medical Centers' inpatient transition of care CCDA.

When a patient was discharged from the hospital, a notification was automatically sent to the PHS Health Information Exchange Hub service ("The Hub"). The Hub was configured to wait 32 hours before calling the CDA Factory to generate a transition of care CCDA for a given patient. The wait time enabled clinicians to finalize the discharge documentation.

Once the transition of care CCDA was generated, it was sent back to The Hub, which in turn called the "Sensitive Data Detection System (SDDS)" to determine whether the CCDA document contained sensitive data or not. If the document did not contain sensitive data, it was stored in a central document repository and automatically sent to the patient's primary care physician and/or referral provider through The Massachusetts Health Information Highway (The Mass HIway) [13]. If the transition of care document was identified as containing sensitive data, the document was stored in the same central document repository, but flagged as "sensitive document" and not routed to The Mass HIway. See ▶ Figure 1 for a diagram of the system.

Other PHS hospitals generate transition of care CCDAs using their respective vendor EHRs. Vendor EHRs have the option to use The Hub to send their CCDAs to the Mass HIway, or they can send the documents directly. EHRs that utilize The Hub have their CCDAs checked for sensitive data using SDDS and following the process outlined above.

2. Objectives

The objective of this paper is to describe the implementation of the "Sensitive Data Detection System (SDDS)," a rule-based application used to screen CCDAs generated when patients are discharged from PHS hospitals. We also report the results of SDDS use across an 18 month period.

3. Methods

The SDDS was implemented using the "Enterprise Clinical Rules System" (ECRS), a web-based decision support service previously described elsewhere [14]. In order to process and translate data from different EHRs, ECRS is implemented using a common "Patient Information Model" (PIM) [14]. The PIM is a "normalized" data model used by all the rules contained in ECRS, including the SDDS rules.

Once ECRS receives a request from The Hub, it calls another service, known as the "Patient Factory". The Patient Factory service is responsible for parsing and mapping the CCDA content into the PIM. The Patient Factory parses not only the structured data of the CCDAs (e.g., medications, problems), but also the corresponding narrative text (human-readable part of the CCDA) associated with each structured entry. After parsing and mapping the CCDA data to the PIM, the service is also responsible for classifying the structured data.

The classification process relies on calls to specific classification services for medications, allergies, and problems. Problems present in the CCDA that are represented using SNOMED CT codes [15] are classified using locally defined classes. For example, code "442537007" corresponding to "Non-Hodgkin lymphoma associated with Human immunodeficiency virus infection" is classified as belonging to the "HIV Positive" class. Medications and allergies represented using RxNorm codes [16] are classified using the "Enhanced Therapeutic Classification" (ETC) [17]. For example, code "317150" corresponding to "Ritonavir 100 MG Oral Capsule" is classified as "Antivirals," "Antiretrovirals," "Antiretroviral – Protease Inhibitors," and "Anti-Infective Agents." Once the data classification is completed, the results are added to the PIM. The Patient Factory returns the classified PIM instance to ECRS and the SDDS rules are executed.

The requirements for the rules were defined by the PHS Health Information Management team and a group of subject matter experts (clinicians that work in the federally funded substance abuse treatment clinics). The rules were based on pre-existing rules used in the manual process for the release of documents and further refined in meetings with the group of subject matter experts.

The SDDS rules were built to detect three categories of sensitive data: (a) data suggesting that the patient is HIV positive, (b) data suggesting that the patient is being treated for substance abuse in a federally funded program, and (c) patients being tested for HIV (independent if the result was positive or negative – state law requirement). The data types evaluated by the SDDS rules are summarized in ▶Table 1.

The rules were organized into groups to facilitate maintenance and execution. The first group consists of the "classification rules" that apply to all data types and are responsible for determining the basic "Clinical States" of the patient. Clinical States are the results of inferences applied to the patient data. Different data types can be used to achieve the same "Clinical State", such as a problem in the problem list or the result of laboratory test can indicate the presence of a disease or a state. Also, basic Clinical States can be further refined into higher levels of inference called "Final Clinical States". Once all Final Clinical States are created, the last group of rules just determines if the Final Clinical States indicative of sensitive data are present or not. For example, the Final Clinical State of a patient being "HIV positive" can be achieved by a problem in the problem list, a coded antiretroviral prescription, or a recorded allergy to antiretroviral medication. Examples of the rules can be found in ▶Table 2.

Rules to detect allergies were added to identify previous treatments with the drugs of interest. Rules to detect sensitive laboratory results were implemented in a second phase. Initially they were not needed since the CDA Factory automatically excludes sensitive laboratory results from being included in the transition of care CCDAs. Once the rules started analyzing the CCDAs produced by a

vendor EHRs that did not filter out sensitive results from the CCDA document, the rules became necessary and were added.

Besides searching for coded data, a string search in the narrative text was implemented because not all medications and/or allergies can be represented using the RxNorm codes allowed by CEHRT. Since there is no restriction in the HL7 CCDA standard about what should or not be included in the narrative text, the complete text parsed by the Patient Factory includes all the information that was included in the human readable part, including eventual characters used as field delimiters and format markup.

After the implementation of the SDDS rules, the results were evaluated during two different time periods. The first evaluation period was soon after the rules implementation. During this evaluation, all CCDAs identified as containing sensitive data were manually reviewed. The second evaluation period was after the rules were further refined to reduce the false positives identified during the first period and new rules were added to detect laboratory results. In the second evaluation period, in addition to reviewing all documents identified as containing sensitive data, two samples of CCDAs not identified as containing sensitive data were also reviewed.

4. Results

The SDDS has been in production use since January 2015. The initial implementation of SDDS was based on a set of 182 rules. This set included two rules for identifying problem subsets, two rules for identifying medication classes, two rules for identifying allergy classes, 88 rules for identifying medications in narrative text, and 88 rules for identifying allergies in narrative text. This initial implementation took four months total, taking into account analysis of requirements, rule implementation and testing, implementation of new classification service for medications, and production deployment.

We first evaluated SDDS after its production release with the initial set of 182 rules. This first evaluation was designed to verify if screening rules were correctly identifying documents containing sensitive data. During a period of 16 days the SDDS was activated 6,020 times, corresponding to the same number of transition of care CCDAs analyzed. A total of 5,841 (97%) CCDAs were identified as not containing sensitive data, while 179 (3%) were identified as containing sensitive data. The 179 documents were manually reviewed by the first author and 79 (44%) were considered "true-positives", indeed containing sensitive data, and 100 (56%) were considered "false-positives". The presence of antiretroviral medications without HIV in the problem list was considered to be a true-positive for the purpose of screening, since it may be indicative of the disease even if it is not documented in the problem list. The details of the true-positive results can be found in ▶Table 2.

All the false-positive results were caused by the narrative text search related to abbreviations of medications and allergies. The string search process found parts of commonly used words or other drugs. The details of the false-positive results can be found in ▶Table 3.

Taking into account these initial results, the SDDS rules were refined with the objective of reducing false positives. All searched strings that generated false positives were reviewed by subject matter experts to determine if they could be improved. The reviewers concluded that the rule using the "DDI" abbreviation should not be used, since the drug "Didanosine" is never prescribed by itself. In addition, new rules were created to analyze laboratory results. The revised and expanded set of rules included 311 rules total: two rules for identifying problem subsets, four rules for identifying medication classes, and 14 rules for identifying medication product codes that cannot be classified, two rules for identifying allergy classes, 100 rules for medications in narrative text, 87 rules for the allergies in narrative text, and 102 rules for identifying specific laboratory results. It took just a few days to implement and test the new rules and have the enhancements available in production.

After these enhancements the results were analyzed for a 30 day period. During this second analysis period, the SDDS was activated 6,935 times. The average time to process a single CCDA including parsing the document, classifying the codes, and executing the rules was 1.66 seconds (median 1.07 sec, minimum 0.19 sec, maximum 13.66 sec). Of all analyzed CCDAs, a total of 6,623 (95.5%) were identified as not containing sensitive data, and 312 (4.5%) were identified as containing sensitive data. All documents containing sensitive data were manually reviewed by the first auth-

or and 307 (98.4%) were considered "true-positives", indeed containing sensitive data, and 5 (1.6%) were considered "false-positives". The details of the true-positive results can be found in ▶Table 4.

In addition, two samples of 50 CCDA documents were also manually reviewed by the first author to confirm the existence of false negatives. Each of the samples was randomly obtained from 2 separate collections of CCDAs considered as having a high potential of containing sensitive data, but not identified by the SDDS rules. The first collection was created for checking the "HIV positive" and "HIV testing" categories (e.g., the collection contained CCDAs that had the term "HIV" or other terms indicative of HIV results in any CCDA section). The review of this first sample identified one false-negative document for the category HIV positive (2%) and ten false-negatives in the HIV testing category (20%). The false-negative for the "HIV positive" category occurred because the patient had no structured entries for problems, medications, or allergies in the CCDA. The false-negatives for the "HIV testing" category were caused by laboratory results mapped to an incorrect LOINC code.

The second collection of documents was created for the detection of "Substance abuse" treatment in a federally funded program category (e.g., documents that contained terms such as "drug abuse" or "alcohol abuse" in any CCDA section). The review of this second sample identified five false-negatives (10%). The false-negatives for the "Substance abuse" category were caused by free text comments in the discharge instructions or in the hospital course summary indicating that the patient was being treated in a federally funded program (e.g., "please do not miss your scheduled appointment at clinic 'x'").

5. Discussion

Meaningful Use rules require CCDA documents be sent electronically at transitions of care. However, federal and state laws prohibit the transmission of documents containing sensitive data unless a patient consent is obtained for each transmission. The SDDS was implemented to identify CCDAs that might contain sensitive data and should not be transmitted outside our organization without specific patient consent.

The results of our initial evaluation confirm that a rule-based method can efficiently identify CCDAs containing sensitive data, particularly considering the large number of documents automatically generated and transmitted. The proportion of documents identified during the second evaluation as containing sensitive data (312, 4.5%) is considered manageable for subsequent manual release to the patient or an authorization request. The initial relatively high proportion of false-positives (100 CCDAs or 1.7% of all documents analyzed) seemed appropriate for an initial implementation of a new screening tool, given the intent of avoiding missing the identification of documents that contain sensitive information (false-negatives), but also the opportunity to quickly release incorrectly identified documents during the manual review. The subsequent enhancements to the SDDS rules decreased significantly the number of false-positives (from 56% to 1.6%) without compromising detection.

The results clearly confirm the importance of processing non-structured (narrative) data in a CCDA. The narrative text associated with CCDA entries contains information that frequently does not exist elsewhere in the structured parts of the document. Our second period results demonstrated that a simple string search process was able to identify 133 CCDAs that would otherwise have not been identified. However, narrative data might also contain information (e.g., comments) that can easily mislead string-matching methods, such as those offered within ECRS. As demonstrated in the examples provided in ▶Table 3, some search strings were present in commonly used words, resulting in confirmed false-positives. Just the removal of the "DDI" abbreviation reduced the false-positives by 87%, without affecting the number of true-positives. We are also considering the adoption of a more robust Natural Language Processing (NLP) solution for the CCDA narrative text analysis.

Conversely, these results also confirm that traditional screening of commonly used structured data types (e.g., problems, medications) is not sufficient by itself to detect sensitive data. For example, of the 74 CCDAs identified in the second evaluation period as containing HIV data, only 2 (2.7%) were identified solely by coded problems and 25 (33.8%) were identified just by coded medi-

cations or allergies. Allergies, a data type that is not commonly used for screening, showed promising results: 22 (29.7%) CCDAs in the HIV category were identified only by allergies.

The implementation of the SDDS was straightforward, mostly because of the infrastructure available and operational at PHS. The existing services to parse CCDA documents (Patient Factory) and to execute production rules (ECRS) provided the key components of SDDS. The only new service implemented for this project was the RxNorm to ETC classification service. The subsequent addition of the laboratory results did not require any additional Patient Factory work given the availability of multiple data types parsers for CCDA documents. However, additional SDDS extensions might require new Patient Factory parsers depending if the data types of interest are already in use or not. Similarly, the expansion of the SDDS rules was easily implemented due to the structure of the rules. The use of "clinical states" allowed easy addition and removal of classification rules without having to restructure subsequent rules that depend on the clinical states existence.

During the second analysis period we were able to demonstrate a very low number of false-negatives for HIV positive patients (2%). These results confirm that a limited number of data types are sufficient to effectively screen CCDAs for HIV positive data. However, the second analysis also identified an issue with the assignment of the appropriate LOINC codes to laboratory results at one of the hospitals. As expected, the SDDS rules are directly affected by incorrect or missing CCDA data. We are evaluating the implementation of additional rules to help identify and handle data quality issues.

The review of the CCDAs in the treatment for substance abuse category showed an already expected higher false-negative rate (10%) since it was not the intent of the rules to find patients with substance abuse, but only patients treated in a federally funded program. Since the information if the patient is being treated in an outpatient federally funded program is not readily available in a CCDA created for an inpatient visit, we had to use indirect mechanisms to identify these patients, such as medications that are commonly prescribed at federally funded substance abuse clinics. In order to reduce the number of false-negatives, we could have added rules to withhold CCDAs with substance abuse codes in the problem list, but such approach would have caused a large number of false-positives. If we were to detect problems that indicated substance abuse, we would have filtered all the false-negative CCDAs. However, an additional 27 CCDAs (54%) would have been filtered for patients that had no indication of being treated in a federally funded program and that would benefit from the continuity of care information.

The identification of documents containing sensitive clinical information for regulatory purposes appears to be a novel application for production rules systems. We performed a literature search and were not able to find published studies of similar systems. Most of the published literature using similar techniques addresses the de-identification of clinical documents [18, 19].

A limitation on the generalization of this work is the use of a sophisticated decision support platform, which includes layers for reference terminologies, data models, interoperability standards, and production rules. The rule application described in this study required software engineers to extend the underlying platform, and informaticians to create and maintain the content – resources that may not be available in other institutions. However, given the remote capabilities of the ECRS, other institutions could make use of a remote decision support service for screening sensitive data. Future research would be needed to determine the feasibility of such remote services, considering the variability of how data are structured inside CCDA documents and state-specific regulations.

Other limitations are associated with the specific portions of CCDA documents that are being analyzed, since other sections may also contain sensitive data. Similarly, we used relatively simple string search methods being used, which are not able to always discriminate abbreviations from sequences of characters present in ordinary words. Despite these limitations, the results demonstrated that it still possible to detect sensitive data with a limited data set, resulting in a high number of true positives and a low false negative rate.

We have been trying to achieve a balance between protecting patient information and minimizing disruptions during care transitions. We believe that with the above described process we were able to protect the patient information when necessary while at the same time allowing for a prompt transition of care. At the beginning of this year, the government has proposed the revision of title 42 of the Code of Federal Regulations part 2 [20]. This revision might facilitate the electronic exchange of documents containing sensitive information, allowing a more seamless continuation of care.

6. Conclusion

The analysis of the "Sensitive Data Detection System" (SDDS) demonstrates that production rules can be used to automatically identify documents containing sensitive data. The SDDS was relatively simple, largely because of the decision support infrastructure already available at PHS. Similar document screening systems may not be easily implemented at other institutions, considering the various components necessary to efficiently process different data types. Future work includes additional analyses to determine other data types that can further improve the screening process, as well as performance comparisons with screening process available within commercial EHR systems.

Questions

The most challenging aspect in automatically analyzing an electronic document for sensitive patient data is:

- A. Availability of a software platform to represent production rules
- B. Complexity of the underlying document interchange standard
- C. Quality and consistency of the available structured data
- D. Complexity of the regulations pertaining to sensitive data

Correct answer: C

Clinical Relevance Statement

The Centers for Medicare & Medicaid Services' Stage 2 final rule requires that hospitals provide transitions of care documents in order to qualify for "meaningful use" incentive payments. State and federal laws prohibit the transmission of documents containing sensitive data. We implemented a methodology that enables a balance between the need to protect patient information and the need to support data exchange during care transitions.

Conflicts of Interest

HSG is on the Scientific Advisory Committee for Clearsense in Jacksonville, FL and is a consultant for PSMI in Point Richmond, CA. The other authors declare that they have no conflicts of interest in the research.

Protection of Human Subjects

The study was submitted to the Institutional Review Board and was considered to be a Clinical Quality Improvement/Measurement that does not need IRB review.

Acknowledgments

The authors would like to acknowledge the assistance of the PHS Health Information Management team for providing the requirements for the rules and the PHS Knowledge Management team and PHS Medication decision support services team for providing content and support for the classification services.

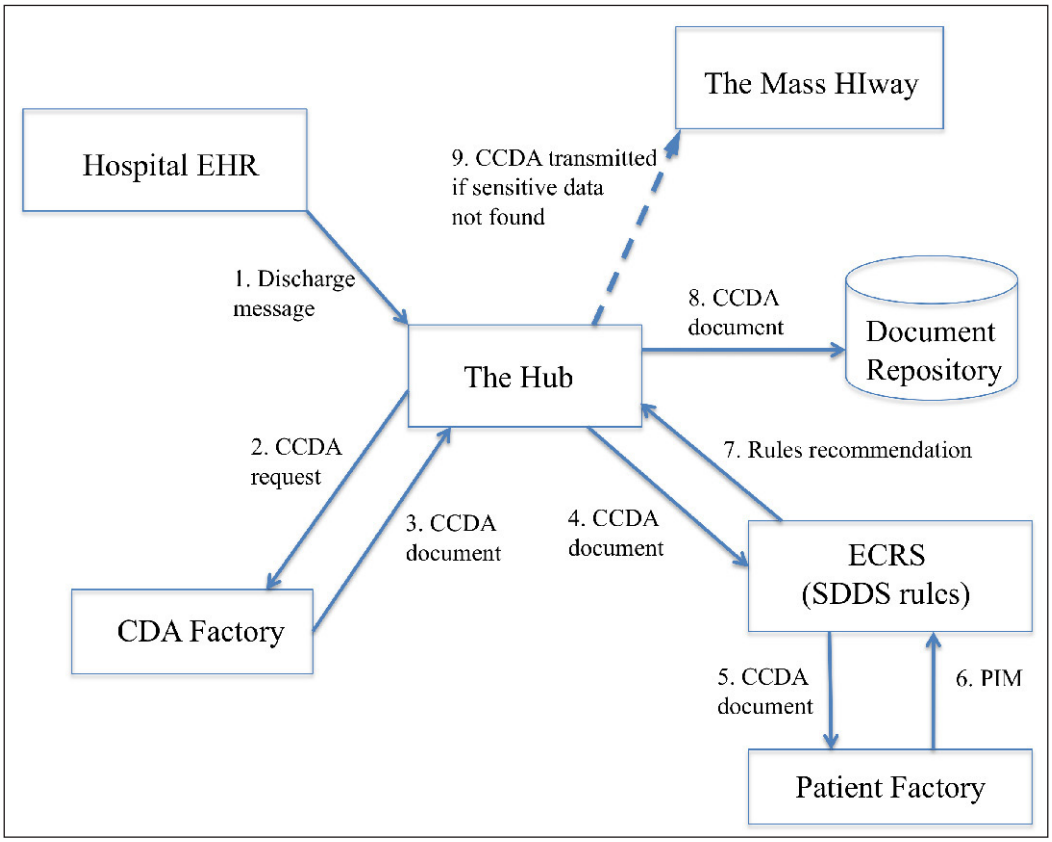

Fig. 1 Diagram of the complete system structure. The numbers indicate the data flow sequence.

Table 1 Data types evaluated by the SDDS rules

Data type	Evaluation description
Problems	Presence of individual codes and/or classes (for HIV only)
Medications	Presence of the medication classes of interest, or drug name and/or its common abbreviations in the narrative text (string search)
Allergies	Presence of the drug allergy classes of interest, or drug name and/or its common abbreviations in the narrative text (string search)
Laboratory Results	Presence of individual codes independent of the result (for HIV only)

Table 2 Example of rules created to detect sensitive data

Rule type	Rule description
Classification Rule	*if* **thePatient** has active free-text Medication "ABACAVIR" *then* add Active Clinical State with code **ANTIRETROVIRAL** and code system is SNOMED, qualifier name **RECORD_OF**, qualifier name code system is SNOMED, qualifier value **ADMINISTRATION_OF_MEDICATION** and qualifier value code system is SNOMED to **thePatient**;
Intermediate States	*if* **thePatient** has Active Clinical State with code **ANTIRETROVIRAL**, code system is SNOMED , qualifier name **RECORD_OF**, qualifier name code system is SNOMED, qualifier value **ADMINISTRATION_OF_MEDICATION** and qualifier value code system is SNOMED *then* add Active Clinical State with code **HIV_POSITIVE** and code system is SNOMED to **thePatient**;
Final Clinical State	*if* **thePatient** has Active Clinical State with code **HIV_POSITIVE** and code system is SNOMED *then* add Observation Request with type **REASON**, observation code **HIV_POSITIVE**, alternative text "**HIV positive**" to the request list of **theAction**; add Active Clinical State with code **RESTRICTED** code system is SNOMED, qualifier name **RECOMMENDATION**, qualifier name code system is SNOMED, qualifier value **ISSUED**, qualifier value code system is SNOMED to **thePatient**;

Table 3 True-positive results with frequency of occurrence by type of data (First analysis period with initial set of 182 rules)

Sensitive Data Categories	Types of data that identified the sensitive data	Identified only by free-text search	Identified only by coded data	Identified by coded and free-text data	Total number of true positives
HIV Positive	Medications only	2	0	16	18
	Allergies only	0	0	23	23
	Problems only	N/A	3	N/A	3
	Combinations of medications, allergies, and/or problems	0	0	17	17
				Total identified	**61**
Treatment for Substance abuse in a federally funded program	Allergies only	1	0	1	2
	Medications only	15	0	1	16
				Total identified	**18**

Table 4 False-positive results identified by cause (First analysis period with initial set of 182 rules)

Free-text search that caused False Positive (FP)	Number of FP			Examples of the text matched
	Allergies	Medications	Total	
"ABC" as an abbreviation for "Abacavir"	1	0	1	Last name of a provider
"AZT" as an abbreviation for "Zidovudine"	2	6	8	Aztreonam
"DDI" as an abbreviation for "Didanosine"	9	78	87	addition, additionally, last name of provider
"ENF" as an abbreviation for "Enfuvirtide"	4	0	4	Last name of a provider

Table 5 True-positive results with frequency of occurrence by type of data (Second analysis period with revised set of 311 rules)

Sensitive Data Categories	Types of data that identified the sensitive data	Identified only by free-text search	Identified only by coded data	Identified by coded and free-text data	Total number of true positives
HIV Positive	Allergies only	0	22	0	22
	Medications only	19	3	1	23
	Problems only	N/A	2	N/A	2
	Combinations of medications, allergies, and/or problems	0	3	24	27
				Total identified	**74**
Treatment for Substance abuse	Allergies only	0	0	0	N/A
	Medications only	113	24	57	194
	Combinations of medications and allergies	1	0	2	3
				Total identified	**197**
HIV testing	Laboratory results only	N/A	36	N/A	36

References

1. Health Information Technology: Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, 2014 Edition; Revisions to the Permanent Certification Program for Health Information Technology. Federal Register 2012: 77(171); 45 CFR Part 170.
2. Medicare and Medicaid Programs; Electronic Health Record Incentive Program – Stage 2. Federal Register 2012: 77(171); 42 CFR Parts 412, 413, and 495.
3. Health Level Seven International. HL7 Implementation Guide for CDA® Release 2: IHE Health Story Consolidation, DSTU Release 1.1 (US Realm) Draft Standard for Trial Use July 2012.
4. Massachusetts Laws: HIV test; informed consent; disclosure of results or identity of subject of test. https://malegislature.gov/Laws/GeneralLaws/PartI/TitleXVI/Chapter111/Section70F . Last accessed on June 1st, 2016.
5. Electronic Code of Federal Regulations [Internet]. Available from: http://www.ecfr.gov/cgi-bin/text-idx?rgn=div5;node=42%3A1.0.1.1.2 . Last accessed on June 1st, 2016.
6. Cohen GR, Adler-Milstein J. Meaningful use care coordination criteria: Perceived barriers and benefits among primary care providers. J Am Med Inform Assoc 2016; 23(e1): e146-e151.
7. Rothstein MA, Talbot MK. Compelled Disclosure of Health Information. Protecting against the Greatest Potential Threat to Privacy. JAMA 2006: 295(24): 2882-2885.
8. Rothstein MA. The Hippocratic Bargain and health Information Technology. J Law Med Ethics 2010: 38(1): 7–13.
9. Sittig DF, Singh H. Rights and responsibilities of users of electronic health records. CMAJ 2013M; 184(13): 1479-1483.
10. Health Level Seven International. HL7 Clinical Document Architecture, Release 2.0. Normative Edition – May, 2005.
11. Health Level Seven International. HL7 Implementation Guide: CDA Release 2 – Continuity of Care Document (CCD) April 2007.
12. Healthcare Information Technology Standards Panel. HITSP Summary Documents Using HL7 Continuity of Care Document (CCD) Component HITSP/C32. 2009 July.
13. The Massachusetts Health Information Highway. http://www.mass.gov/eohhs/gov/commissions-and-initiatives/masshiway/. Last accessed on June 1st, 2016.
14. Goldberg HS, Paterno MD, Rocha BH, Schaeffer M, Wright A, Erickson JL, Middleton B. A highly scalable, interoperable clinical decision support service. JAMIA 2014: 21(e1): e55-e62.
15. SNOMED CT. Available from: http://ihtsdo.org/snomed-ct/. Last accessed on June 1st, 2016.
16. RxNorm. Available from: http://www.nlm.nih.gov/research/umls/rxnorm/. Last accessed on June 1st, 2016.
17. FDB Enhanced Therapeutic Classification System. Available from: http://www.fdbhealth.com/fdb-medknowledge-foundations/. Last accessed on June 1st, 2016.
18. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. Available from: http://www.biomedcentral.com/1471–2288/10/70. Last accessed on November 7th, 2016.
19. Stubbs A, Kotfila C, Uzuner O. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. Journal of Biomedical Informatics 2015; 58: S11–S19.
20. Confidentiality of Substance Use Disorder Patient Records. Available from: https://www.regulations.gov/#!documentDetail;D=HHS-OS-2016–0005–0001. Last accessed on June 1st, 2016.