Searching the PDF Haystack: Automated Knowledge Discovery in Scanned EHR Documents

Alexander L. Kostrinsky-Thomas[1] Fuki M. Hisama[2] Thomas H. Payne[3]

[1] College of Osteopathic Medicine, Pacific Northwest University of Health Sciences, 200 University Pkwy Yakima, Washington, United States
[2] Division of Medical Genetics, Department of Medicine, University of Washington School of Medicine, Seattle, Washington, United States
[3] Department of Medicine, University of Washington School of Medicine, Seattle, Washington, United States

Address for correspondence Alexander L. Kostrinsky-Thomas, BA, 6223 39th AVE NE, Seattle, Washington, 98115, United States (e-mail: akostrinskythomas@pnwu.edu).

Appl Clin Inform 2021;12:245–250.

Abstract

Background Clinicians express concern that they may be unaware of important information contained in voluminous scanned and other outside documents contained in electronic health records (EHRs). An example is "unrecognized EHR risk factor information," defined as risk factors for heritable cancer that exist within a patient's EHR but are not known by current treating providers. In a related study using manual EHR chart review, we found that half of the women whose EHR contained risk factor information meet criteria for further genetic risk evaluation for heritable forms of breast and ovarian cancer. They were not referred for genetic counseling.

Objectives The purpose of this study was to compare the use of automated methods (optical character recognition with natural language processing) versus human review in their ability to identify risk factors for heritable breast and ovarian cancer within EHR scanned documents.

Methods We evaluated the accuracy of the chart review by comparing our criterion standard (physician chart review) versus an automated method involving Amazon's Textract service (Amazon.com, Seattle, Washington, United States), a clinical language annotation modeling and processing toolkit (CLAMP) (Center for Computational Biomedicine at The University of Texas Health Science, Houston, Texas, United States), and a custom-written Java application.

Results We found that automated methods identified most cancer risk factor information that would otherwise require clinician manual review and therefore is at risk of being missed.

Conclusion The use of automated methods for identification of heritable risk factors within EHRs may provide an accurate yet rapid review of patients' past medical histories. These methods could be further strengthened via improved analysis of handwritten notes, tables, and colloquial phrases.

Keywords

- electronic health records
- portable document format
- optical character recognition
- natural language processing
- machine learning
- evaluation

received
December 6, 2020
accepted after revision
February 1, 2021

Georg Thieme Verlag KG,
Rüdigerstraße 14,
70469 Stuttgart, Germany

DOI https://doi.org/10.1055/s-0041-1726103.
ISSN 1869-0327.

Background and Significance

Most electronic health records (EHRs) contain scanned documents sent to clinicians via fax or other means.[1] Due to these scanned documents not being digitized, they are not readily retrievable, searchable, nor classifiable/organizable.[2] While in theory EHRs allow for seamless electronic health information exchange (HIE), in practice this is not always the case, and as a result documents are often transferred via digital scanning or fax.[3] The content of these scanned documents is not always included in searches conducted by search tools available in commercial EHRs, and therefore the portable document format (PDF) files must be manually reviewed by clinicians.[4] Because there may be many such PDF documents and other outside documents in a patient's EHR, there is risk that clinically important information may not be viewed and acted on by treating clinicians because the quantity of information within these documents can be overwhelming.[5] Automating the search for this information using optical character recognition (OCR) and natural language processing (NLP) through already-existing commercial application programming interfaces, if accurate and complete, might reduce this risk. Though OCR and NLP technologies are broadly applied in other domains, there are no reports of EHRs to extract information from PDFs in general and for cancer risk factors specifically.[6,7]

Objectives

The objective of this study was to demonstrate the feasibility and test the use of automated tools to scan for latent EHR data that could have clinical relevance when determining a patient's risk for heritable breast cancer. The efficacy of these tools was subsequently evaluated via comparison to manual physician review.

Methods

As part of the study of risk factors for heritable cancer in the EHR,[8] we studied records of a random sample of 299 women $\geq$30 years of age who were seen more than five times or hospitalized two or more times at the University of Washington (UW) Medicine health system in Western Washington state between April 2018 and April 2019. For the primary study, the entire EHR was manually reviewed to identify risk factors for heritable breast and ovarian cancer taken from the National Comprehensive Cancer Network (NCCN) Guidelines for Hereditary Testing Criteria, Version 1.2020.[9] The EHR included notes, test results, and demographic information and also outside records and faxes that are scanned and stored in the EHR in PDF format. The focus of the present study is the development and testing of automated methods to identify risk factors for heritable breast and ovarian cancer in the scanned records within the EHR.

To assess the efficacy of an automated system, we first reviewed the same documents manually. Two physicians, both of whom were experienced in risk factor identification, manually reviewed all 91 PDF documents in the outside records section of the EHRs of seven study patients drawn from the records of study subjects. We chose to select seven patients because the 91 pages in their records fit both the requirements for this study as well as the time constraints for the manual deidentification and review processes. All PDF documents and outside records associated with their UW Medicine EHR were printed, manually deidentified with a black pen, and subsequently reviewed for risk factors by the two physicians. When reviewing EHR data for latent information, deidentification would normally not be necessary. However, our automated methods used tools from companies with which we had not yet instituted a Business Associate Agreement; for this reason, we manually removed all identifying data before using their services and verified deidentification by having a second physician review of the same documents again.

To test the ability of an automated system to identify risk factors, the same deidentified PDF documents and outside records were uploaded to Amazon's Textract service (Amazon.com, Seattle, Washington, United States), using an Amazon S3 bucket to securely store the documents prior to analysis. Amazon Textract is a machine learning service that automatically extracts text and other data from scanned documents, and outputs it in a plain text format.[10] This plain text was subsequently entered into CLAMP,[11] a clinical language annotation modeling and processing toolkit, for analysis and classification based upon the entity, relationship, and syntax.

After processing via CLAMP, we used a custom-written Java application to map CLAMP's output to the National Comprehensive Cancer Network Guidelines for referral to genetic counseling.[7] The Java application began by searching the plain text output from CLAMP for a list of keywords taken from the NCCN Guidelines, and then used a negation detection algorithm within CLAMP to exclude entities that represented negative clinical findings.[12] Finally, the application displayed results summarizing all positive clinical entities found, as defined by NCCN Guidelines, for a given document (►Fig. 1).

Results

In 91 printed pages from seven patients' EHRs, 15 risk factors were identified by both physician reviewers across 91 pages; the automated pipeline identified 12 (80%) of these, missing three risk factors total (►Table 1). However, the automated pipeline identified that an additional six risk factors were not identified by at least one of two human reviewers. The data contained in these pages varied widely; they included patient encounter summaries, raw laboratory results, patient completed surveys, and demographic information.

Importantly, we were able to identify scenarios in which our automated screening tool was less accurate than a physician reviewer. Automated methods were less adept than humans in recognizing risk factors that were present in tables, that were handwritten, and that were contained within sentences that did not follow standard English grammatical rules. For example, our screening tool was unable to

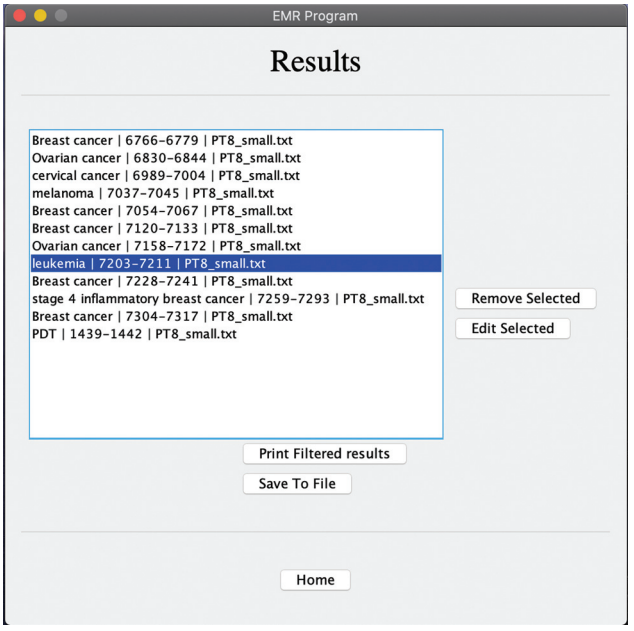

Fig. 1 Screenshot of the Java application, displaying identified risk factors from a sample patient chart.

identify a paternal diagnosis of melanoma given the phrasing, "Immediate family history of Melanoma? Yes father." Further, our program missed a risk factor simply because we had not taught it to identify some proper nouns. For example, Myriad myRisk (Myriad Genetics, Salt Lake City, Utah, United States) is a genetic test designed to evaluate the genes associated with multiple hereditary cancer syndromes. However, since our dictionary within CLAMP did not contain the product name, it was unable to identify and subsequently classify that the patient had received a positive genetic test result.

While an important part of evaluating the automated screening tool is identifying the failures, there were also scenarios in which the tool excelled compared with manual physician review. For example, in patient #2, both physician reviewers missed the phrase "Maternal Grandfather Prostate Ca" that was listed within a family history but was formatted untraditionally due to an unnecessary line break. In patient #6, physician reviewers both missed "Family history of gastric cancer" due to its untraditional location in a pathology result.

In summary, automated review and human review were complementary in finding risk factors; most but not all risk factors found by human review were identified by the automated system, but both automated and human review found risk factors missed by the other (►Table 1).

Discussion

The results of our pilot study show that the potential of existing publicly available software packages linked with our Java program could be combined to create a pipeline to extract important clinical concepts from scanned records contained in the EHR. Key components to this pipeline are Textract and CLAMP which are both broadly available. We found that in this pilot, the pipeline detected most risk factors found by human clinician reviewers in extracting latent data from EHR scanned documents that would otherwise require time-consuming manual review and therefore may be missed. Of the parent study of 25 women that were found to have met NCCN criteria for referral for further genetic risk evaluation, 13 (52%) had no record of a referral to a Medical Genetics clinic; through the notification and subsequent screening of these 13 patients, these data could assist in averting new incidence, morbidity, and mortality from heritable breast and ovarian cancer.

Two areas exist for possible improvement of the pipeline's accuracy. The first is an improvement in the quality of source material. Approaches focusing on this aspect include increasing the quality of printed documents, specifically using higher quality printers, fine-tuning contrast and brightness of individual PDF pages, and finally using higher quality scanners that can process images above 600 dots per inch. While the results from our pilot study were not negatively impacted by low quality source material, the effects are well-documented within contemporary literature.[13] The second approach for quality improvement stems from identifying concepts within the text, including the previously discussed dictionary expansion to help improve medical entity detection. However, further improvement could also be made through the implementation of a post-scan language correction system and in other ways. Previous studies have shown that grammar and spelling correction can be trained and subsequently performed on text containing medical terminology, in a process that reliably decreases identification errors.[14] If accurate, these methods could help reduce the number of false negatives extracted from scanned documents.

As mentioned in the results, our program failed to identify a patient's paternal history of melanoma due to unusual phrasing in a question/answer format. While post-scan language correction might help alleviate some of the errors found in many clinical notes, further advances are required to be made within the NLP field before a higher degree of accuracy can be obtained. Since quantifying the accuracy of an automated algorithm relies upon comparison to manual methods, future work may include implementation of confidence scores. While still an area of ongoing research, these scores, such as those implemented within IBM Watson, incorporate a measure of their confidence within their prediction, allowing users to further evaluate the likelihood of their given result.[15]

Finally, it is important to note that the developed pipeline could be used to extract latent data pertinent to health conditions other than heritable cancer. While identifying risk factors for heritable forms of breast and ovarian cancer was the scope of this study, additional clinically important information also exists in scanned documents. Similar investigations into latent EHR data have identified benefits to extracting cardiovascular data,[1] pulmonary function tests,[16] health maintenance history, immunizations, and other clinical data that may exist unstructured within patient notes.[17] In the current generation of commercial EHRs, this information does

Table 1 Summary of risk factors found in patient records, as identified by automated and manual review

Patient #	Text identified by automated review	Text identified by manual review	Found by manual reviewer 1?	Found by manual reviewer 2?	Why did automated review miss?	Containing document type
1	"Cancer Maternal Aunt"		Y	Y		Encounter Summary
	"Cancer Maternal Grandmother"		Y	Y		Encounter Summary
	"Maternal Grandfather Prostate ca"	Missed	N	N		Encounter Summary
2	"Family History of Breast Cancer"		N	Y		Encounter Summary
	"Family History of Ovarian Cancer"		N	Y		Encounter Summary
	"Breast Cancer Mother"		Y	Y		Encounter Summary
	Missed	"Reflex to myRisk"	Y	Y	Unable to interpret "Reflex to myRisk"	Laboratory Results
	"Ovarian Cancer Mother"		Y	Y		Encounter Summary
	"Breast Cancer Mother"		Y	Y		Encounter Summary
	"Ovarian Cancer Mother"		Y	Y		Encounter Summary
	"Ashkenazi Jewish ancestry"		Y	Y		Encounter Summary
	"Maternal Aunt Cervical Cancer"		Y	Y		Encounter Summary
	"Breast Cancer Paternal Aunt"		Y	Y		Encounter Summary
	"Breast Cancer Maternal Aunt"		Y	Y		Encounter Summary
3	Missed	"Family history of cancer—mother, sister"	N	Y	Handwritten note	Patient Survey
4	"FHx of Breast Cancer and Uterine Cancer"		N	Y		Encounter Summary
	Missed	"Immediate family history of melanoma? Yes father"	N	Y	Sentence structure/ comprehension issue	Encounter Summary
5	Missed	"Cancer: grandfather"	Y	Y	Text was present in table format	Patient Survey
6	"Family History of Gastric Cancer"	Missed	N	N		Operative Report
	"Other Malignant Neoplasm of Skin"		N	Y		Operative Report
7	"History of Malignant Neoplasm of Breast"		Y	Y		Encounter Summary
	"Family History of Malignant Neoplasm of Breast"		Y	Y		Encounter Summary
	Missed	"Mother: colon cancer, breast cancer"	Y	Y	Handwritten note in table format	Patient Survey

not necessarily trigger or satisfy health maintenance reminders, and unless it is manually read and entered, what is contained in these scanned records may not be reflected in the EHR past medical history, patient problem lists, or lists of allergies. As others have noted, the literature devoted to scanned documents and images within EHRs is smaller than we expected given the importance of this commonly used means for HIE in the early decades of EHR use in our country.[18]

Our study is limited by its small size—it is a pilot—and by the population that we used which is from an academic center. The number of cancer risk factors identified in scanned records may be different in other populations. Others who use externally hosted tools may need a Business Associate Agreement to meet HIPAA requirements. Though the automated review did not detect all risk factors in the text, it reduces the risk that busy clinicians would miss

important information in the PDFs because of the time required to review them manually. The accuracy of OCR will also depend on the quality and formatting of the original document, something that the receiving health care organization may have some, but limited ability to improve since documents may be transmitted with suboptimal quality.

Conclusion

Automated methods show promise in reducing the risk that clinicians are unaware of clinically important information in scanned and outside records within the EHR. The methods used in our pilot study can augment existing EHR tools to fully leverage EHR content for cancer risk reduction and other benefits. As patient loads continue to rise, and patient encounter time continues to shrink, programs that help streamline a physician's ability to accurately and efficiently review a patient's past medical history are quickly becoming a vital part of the technological toolkit. However, to ensure optimal accuracy and completeness of automated reviews, future work should focus on improving the detection of data stored in tables, handwritten notes, and clinical narratives that do not follow typical English sentence structures. Subsequent studies should ensure that they have a current Business Associate Agreement in place with each of their partners to expedite processing of documents without the need for redaction, to maintain PHI security and overall HIPAA compliance.

Clinical Relevance Statement

Of the 25 women in our study who met the NCCN criteria for referral, only 12 of these had received further genetic risk evaluation. This situation represents a gap in health care that, using these types of automated tools, could reduce morbidity and mortality from breast and ovarian cancer, and potentially, other inherited cancer syndromes. Our research indicates that screening of latent electronic health record data can be efficiently performed using automated processes that are already commercially available and can find most risk factors identified by a human clinician reviewer. Further refinements may increase sensitivity of automated review.

Multiple Choice Questions

1. What role does natural language processing play when designing an automated pipeline to analyze clinical notes?
   a. It can be used to extract text from image or PDF files.
   b. When used in combination with electronic medical records, it is primarily used to translate medical jargon into layman's terms.
   c. When paired with root cause analysis, it can diagnose a patient's condition.
   d. Through analysis of complete sentences and paragraphs, it can help provide context and meaning for otherwise raw data.

   Correct Answer: Correct answer is option *d, through analysis of complete sentences and paragraphs, it can help provide context and meaning for otherwise raw data.* Natural language processing can be used to give a machine the ability to read, analyze, and understand the complexities of human language.

2. In what ways could implementation of an OCR/NLP pipeline within an EHR help physicians in everyday practice?
   a. By analyzing a patient's full medical history, the pipeline can provide an accurate diagnosis on par with human clinicians.
   b. Through methodical processing of patient documents, the pipeline can provide clinical insight that otherwise might have been overlooked.
   c. Using grammar and sentence structure analysis, the pipeline can assess the medical prowess of previous providers, thereby giving the clinician further insight on whether or not to trust prior diagnoses.
   d. Through consideration of past medical history, family history, and genetics, the pipeline can provide treatment recommendations for cancer patients.

   Correct Answer: Correct answer is *b, through methodical processing of patient documents, the pipeline can provide clinical insight that otherwise might have been overlooked.* Leveraging an OCR/NLP pipeline can aid physicians through automated analysis of documents, leading to potentially less time reviewing charts and more time interacting with patients.

Protection of Human and Animal Subjects

This project was approved by the University of Washington Institutional Review Board.

Conflict of Interest

None declared.

Acknowledgments

Dr. Payne reports grants from Brotman Baty Institute, during the conduct of the study. This work was supported by a Brotman Baty Institute for Precision Medicine Catalytic Collaborations Award. We would like to acknowledge the work and support of Ryan Warren with programming and testing of the Java application, along with integration with CLAMP and Textract.

References

1 Moon S, Liu S, Chen D, et al. Salience of medical concepts of inside clinical texts and outside medical records for referred cardiovascular patients. Journal of Healthcare Informatics Research. 2019; 3:200–219
2 Healthit.gov. What Is HIE? | Healthit.Gov. 2020. Accessed November 25, 2020 at: https://www.healthit.gov/topic/health-it-and-health-information-exchange-basics/what-hie
3 Rudin R, Volk L, Simon S, Bates D. What affects clinicians' usage of health information exchange? Appl Clin Inform 2011;2(03): 250–262
4 Rasmussen LV, Peissig PL, McCarty CA, Starren J. Development of an optical character recognition pipeline for handwritten form fields from an electronic health record. J Am Med Inform Assoc 2012;19(e1):e90–e95

5 Farri O, Pieckiewicz DS, Rahman AS, Adam TJ, Pakhomov SV, Melton GB. A qualitative analysis of EHR clinical document synthesis by clinicians. AMIA Annu Symp Proc 2012;2012:1211–1220
6 Mowery DL, Kawamoto K, Bradshaw R, et al. Determining Onset for Familial Breast and Colorectal Cancer from Family History Comments in the Electronic Health Record. AMIA Jt Summits Transl Sci Proc 2019;2019:173–181
7 Jiang X, McGuinness JE, Sin M, Silverman T, Kukafka R, Crew KD. Identifying women at high risk for breast cancer using data from the electronic health record compared with self-report. JCO Clin Cancer Inform 2019;3:1–8
8 Payne TH, Zhao LP, Le C, et al. Electronic health records contain dispersed risk factor information that could be used to prevent breast and ovarian cancer. J Am Med Inform Assoc 2020;27(09): 1443–1449
9 National Comprehensive Cancer Network. Genetic/Familial High-risk Assessment: Breast, Ovarian, and Pancreatic V.1.2020. Accessed August 18, 2020 at: https://www.nccn.org/professionals/physician_gls/pdf/genetics_screening.pdf
10 Amazon Textract. Amazon Web Services, Inc. Accessed January 16, 2021 at: https://aws.amazon.com/textract/
11 Soysal E, Wang J, Jiang M, et al. CLAMP—a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc 2018;25(03):331–336
12 Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(05): 301–310
13 Holley R. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitization programs. Dlib Mag 2009;15:3–4
14 Hládek D, Staš J, Ondáš S, Juhár J, Kovács L. Learning string distance with smoothing for OCR spelling correction. Multimedia Tools Appl 2016;76(22):24549–24567
15 Ferrucci D, Brown E, Chu-Carroll J, et al. Building Watson: an overview of the DeepQA project. AI Mag 2010;31:59–79
16 Sauer B, Jones B, Globe G, Leng J, Lu C, He T, Teng C, Sullivan P, Zeng Q. Performance of an NLP Tool to extract PFT reports from Structured and Semi-Structured VA data. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016; 4(01):10
17 Liang J, Tsou C, Poddar A. A Novel System for Extractive Clinical Note Summarization. Paper presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; 2019; Minneapolis, MN
18 Goodrum H, Roberts K, Bernstam EV. Automatic classification of scanned electronic health record documents. Int J Med Inform 2020;144:104302