Original Paper

Evaluating the Impact on Clinical Task Efficiency of a Natural Language Processing Algorithm for Searching Medical Documents: Prospective Crossover Study

Eunsoo H Park[1,2*], BMedSci; Hannah I Watson[2*], BMedSci, MBChB, MSc; Felicity V Mehendale[3], MBBS, MS; Alison Q O'Neil[2,4], BSc, MEng, EngD; Clinical Evaluators[5]

[1]Edinburgh Medical School, College of Medicine and Veterinary Medicine, University of Edinburgh, Edinburgh, United Kingdom
[2]Canon Medical Research Europe, Edinburgh, United Kingdom
[3]Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
[4]School of Engineering, University of Edinburgh, Edinburgh, United Kingdom
[5]see Acknowledgements, United Kingdom, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Eunsoo H Park, BMedSci
Edinburgh Medical School
College of Medicine and Veterinary Medicine
University of Edinburgh
The Chancellor's Building
49 Little France Crescent
Edinburgh, EH16 4SB
United Kingdom
Phone: 44 1312426792
Email: e.park-7@sms.ed.ac.uk

Abstract

Background: Information retrieval (IR) from the free text within electronic health records (EHRs) is time consuming and complex. We hypothesize that natural language processing (NLP)–enhanced search functionality for EHRs can make clinical workflows more efficient and reduce cognitive load for clinicians.

Objective: This study aimed to evaluate the efficacy of 3 levels of search functionality (no search, string search, and NLP-enhanced search) in supporting IR for clinical users from the free text of EHR documents in a simulated clinical environment.

Methods: A clinical environment was simulated by uploading 3 sets of patient notes into an EHR research software application and presenting these alongside 3 corresponding IR tasks. Tasks contained a mixture of multiple-choice and free-text questions. A prospective crossover study design was used, for which 3 groups of evaluators were recruited, which comprised doctors (n=19) and medical students (n=16). Evaluators performed the 3 tasks using each of the search functionalities in an order in accordance with their randomly assigned group. The speed and accuracy of task completion were measured and analyzed, and user perceptions of NLP-enhanced search were reviewed in a feedback survey.

Results: NLP-enhanced search facilitated more accurate task completion than both string search (5.14%; *P*=.02) and no search (5.13%; *P*=.08). NLP-enhanced search and string search facilitated similar task speeds, both showing an increase in speed compared to the no search function, by 11.5% (*P*=.008) and 16.0% (*P*=.007) respectively. Overall, 93% of evaluators agreed that NLP-enhanced search would make clinical workflows more efficient than string search, with qualitative feedback reporting that NLP-enhanced search reduced cognitive load.

Conclusions: To the best of our knowledge, this study is the largest evaluation to date of different search functionalities for supporting target clinical users in realistic clinical workflows, with a 3-way prospective crossover study design. NLP-enhanced search improved both accuracy and speed of clinical EHR IR tasks compared to browsing clinical notes without search. NLP-enhanced search improved accuracy and reduced the number of searches required for clinical EHR IR tasks compared to direct search term matching.

(JMIR Med Inform 2022;10(10):e39616) doi: 10.2196/39616

KEYWORDS

clinical decision support; electronic health records; natural language processing; semantic search; clinical informatics

Introduction

Background

The benefits of the transition from storing patient information in paper notes to electronic health records (EHRs) have been a topic of debate among health care professionals [1-4]. Many clinicians have expressed dissatisfaction with their current hospital systems, and EHR use is consistently cited as a contributor to clinician burnout [5-7]. Approximately 40% of doctors' time is spent documenting patient information, with evidence showing that this work burden has increased following EHR implementation [8,9]. However, difficulties in quickly and accurately retrieving relevant information from these documents indicate that this wealth of collected information is often not fully used [10,11]. Navigating EHR documents is challenging owing to the complexity of medical text, which tends to include frequent misspellings, abbreviations, specialty-specific acronyms, and clinical shorthand [12-15]. Time-consuming and inaccurate information gathering from EHRs limits the efficiency of wider clinical workflows [16], with some doctors believing that difficulties in retrieving patient information significantly impact face-to-face patient care [17].

Despite the increasing sophistication of general search engines, there remain relatively limited search options within medical record software. One barrier is the need for patient data to be held securely; therefore, access to computing power and shared resources may be limited. To have clinical utility, search facilities must be fast and intuitive for use by time-pressured clinicians, including relatively junior members of staff to whom the task of searching through complex notes is frequently delegated. In addition, the search must handle high variability of text expression as mentioned above. Clinical text is error prone; unlike journals and other publications, there is no editorial control to check for errors. Medical terminology, acronyms, and abbreviations vary between regions and hospitals and even across different specialties; for instance, "CHD" may be related to chronic heart disease (cardiology), congenital heart disease (pediatrics), or congenital hip dislocation (orthopedics). Since clinical care is a high-stakes environment, failure to find relevant information potentially has great implications; to effectively save the time of clinicians, search tools should ideally go beyond document-level results to locate and highlight all relevant sentences or even words within a document. Efforts to achieve easier information retrieval (IR) have included the integration of string search in some EHRs, similar to the "Ctrl-F" or "Find" function that is now frequently available on everyday platforms [18]. However, the effectiveness of string search is limited for heterogeneous clinical text; therefore, studies have also considered semantic search algorithms [19-22]. A large-scale retrospective analysis of searches performed in an EHR found that the use of search varied considerably across and within user roles, with physicians and pharmacists being the most active user groups [19]. A review of the use of search within EHRs found that few articles focused on the impact of search within clinical workflows [23]; one study with 7 diabetes experts found that content-based search was both faster and more accurate than conventional search for finding relevant information [20], another study with 10 family and internal medicine physicians found that semantic search allowed for faster medical notes navigation for IR tasks [21], and a final study with 4 students found that a semantic search tool enabled faster clinical note summarization [22]. Only one of the described studies [20] used a crossover study design. In this paper, a larger study is reported (n=35 valid task completions, n=42 qualitative responses), in which a 3-way prospective crossover study was conducted, comparing a standard string search with no search and with a natural language processing (NLP)–enhanced search. The custom NLP-enhanced search tool combines ontologies with fuzzy matching to offer search functionality, which captures not only semantically related terms (eg, synonyms and hyponyms) but also linguistic alternative spellings and misspellings and word forms of the search term. A simulated clinical environment was used alongside target user feedback to determine whether search tools could make clinical workflows more efficient and reduce clinicians' cognitive burdens when attempting to find information.

Aims and Hypotheses

This study aimed to quantitatively and qualitatively compare the efficacy of 3 search functionalities for IR from medical free-text documents, in terms of their accuracy, speed, and ease of search.

We hypothesized that search tools will allow clinical users to perform simulated clinical IR tasks faster and more accurately than when using no search, with the use of NLP techniques enabling NLP-enhanced search to perform more effectively than string search.

Methods

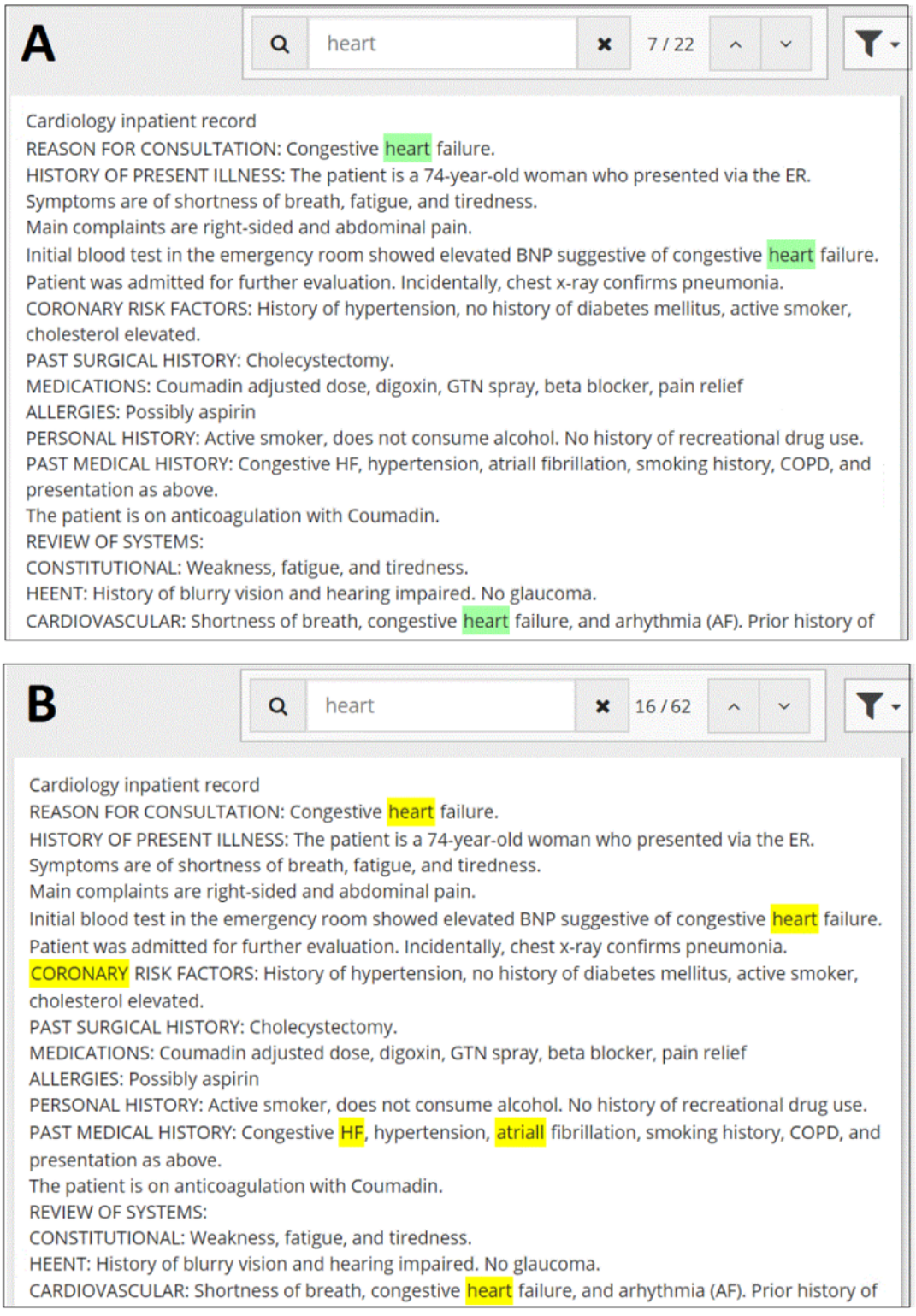
Search Tools

The string search function is an open source JavaScript library implementation [24]. NLP-enhanced search is a proprietary rule-based algorithm (developed at Canon Medical Research Europe) that leverages NLP techniques such as edit distance and stemming in conjunction with medical knowledge bases, notably the Unified Medical Language System semantic network, Metathesaurus [25], and medical abbreviation lists on Wikipedia [26] and OpenMD [27]. These sources are used to expand the original search term into a list of equivalent terms, which are then located in the text. The tool was designed to locate linguistic variants such as misspellings and alternative spellings, word forms, and abbreviations, as well as additional semantic synonyms.

Search tools were integrated into a patient-centric viewer (EHR research software), which allowed the user to type in a search

term and view the highlighted findings within the retrieved subset of documents, which the user could scroll through. In the case of no search, the user was expected to scroll through the patient's EHR to find the relevant information. Figure 1 illustrates the difference between the two search tools in the patient-centric viewer.

Figure 1. Example results for (A) string search and (B) NLP-enhanced search for the search term "heart." String search returned only direct matches to "heart" (green highlights) whereas NLP-enhanced search also returns semantically related terms (yellow highlights) such as the following: "coronary," the misspelling of atrial (fibrillation) as "atriall," and the appearance of "heart" within the abbreviation of heart failure, "HF." NLP: natural language processing.



Simulating a Clinical Workflow

Overview

Free-text medical documents were synthesized for 3 fictional patients. These materials were paired with corresponding sets of 10 IR questions for each patient, grounded in relevant and realistic clinical scenarios. Patient documents were uploaded into the patient-centric viewer. Questions were uploaded onto a custom evaluation platform built using Vue.js, which also displayed the clinical scenarios and task-specific instructions for the evaluator. Below, we describe the document synthesis and question generation in more detail.

Patient Document Synthesis

Three patient profiles were created with varying age, sex, ethnic background, social history, and medical history. The 3 patients were assigned primary medical specialties of respiratory, neurology, and oncology. For each patient, 20 documents were created by selecting and augmenting publicly available anonymized medical documents [28], as well as manually synthesizing additional documents to provide a patient EHR with a coherent chronological sequence of clinical events. Documents were varied and included discharge letters, outpatient clinic letters, operation notes, and general practice referral letters. To imitate real-world medical text, common misspellings, abbreviations, and acronyms were included in the text, using investigator clinical experience (author HW) and reference papers [13].

Clinical Scenarios and Question Generation

For each task, clinical scenarios were designed to focus on real-world situations where information can be extracted from patient notes. To ensure that the tasks were comparable across patients (and therefore interventions), a master template of 10 questions prompting IR was created, which was then tailored to fit each patient scenario. Questions were inspired by those in past medical examinations [29] and investigators' (HW and FM) clinical experience. Requested information resembled that required in typical clinical workflows to support clinical decision-making. Care was taken to ensure that task questions tested the search function and not clinical knowledge or judgement; therefore, all answers could be found by searching the respective patient's notes. Questions required a mixture of multiple-choice and free-text responses. Examples of scenarios and corresponding questions for each patient can be seen in Table 1.

Table 1. Examples of clinical scenarios for each patient and their corresponding question-answer options. Scenarios aimed to simulate a standard clinical workflow, providing context for the questions.

Patient	Example clinical scenario	Example question	Answer type

1	You're worried this may be an exacerbation of a previously present infection. After sending the patient for a chest X-ray and taking bloods, you continue to search for more information.	Does this patient have a history of respiratory infection during the months December 2020-February 2021?	• Select one of the following:
			  • Yes
			  • No
			  • Information not available

		Why was the patient's nitrofurantoin stopped?	• Free text

2	Patient presents to the Emergency Department with confusion and acute stroke-like symptoms. His son reports 2 previous "mini-strokes". You are an ED registrar and send him for a CT head, as per protocol. While waiting for the results you search his history for other contraindications to thrombolysis treatment.	Does the patient have a history of head trauma or stroke between November 2020 and February 2021 (inclusive)?	• Select one of the following:
			  • Yes
			  • No
			  • Information not available

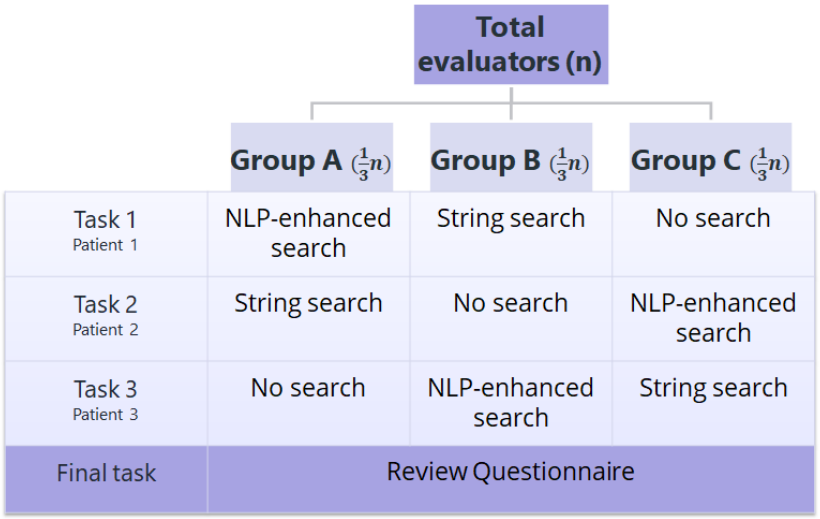
		Search the notes to find the dates of the aforementioned "mini-strokes" (e.g. transient ischaemic attacks).	• Free text
3	You are the new oncologist at the clinic seeing this patient for review. Prior to the appointment you want to check her history by accessing her notes so you can adequately prepare yourself for the consultation.	What is the patient's cancer diagnosis?	• Free text

		Does this patient have a history of any of the following conditions?	• Select all that apply:
			  • Metastases
			  • Hypertension
			  • Epilepsy
			  • Asthma
			  • None of the above

Study Design

The clinical evaluation pipeline was structured as having a prospective crossover trial design; we have illustrated this in Figure 2. Evaluators were banded on the basis of their level of clinical experience before being assigned pseudonymized evaluator IDs that were used for the remainder of the study and analysis. Evaluators in each band were then randomly allocated across the 3 study groups using a random number generator. This yielded 3 groups stratified for level of clinical experience. Each group had a predetermined order of search functionality; once the 3 tasks were completed using the allocated search order, evaluators were asked to fill out a feedback survey that focused on their user experience.

Figure 2. Study design. The 3 tasks were performed using a prospective crossover design, in which each group undertook the tasks in the same order with a predetermined order of the search intervention; the order was different for different groups. Finally, all evaluators were asked to fill in a review questionnaire. NLP: natural language processing.



Evaluator Recruitment and Training

Recruitment for the study was accomplished via professional contacts and advertising on social media channels to reach evaluators from a variety of clinical specialties and years of clinical experience.

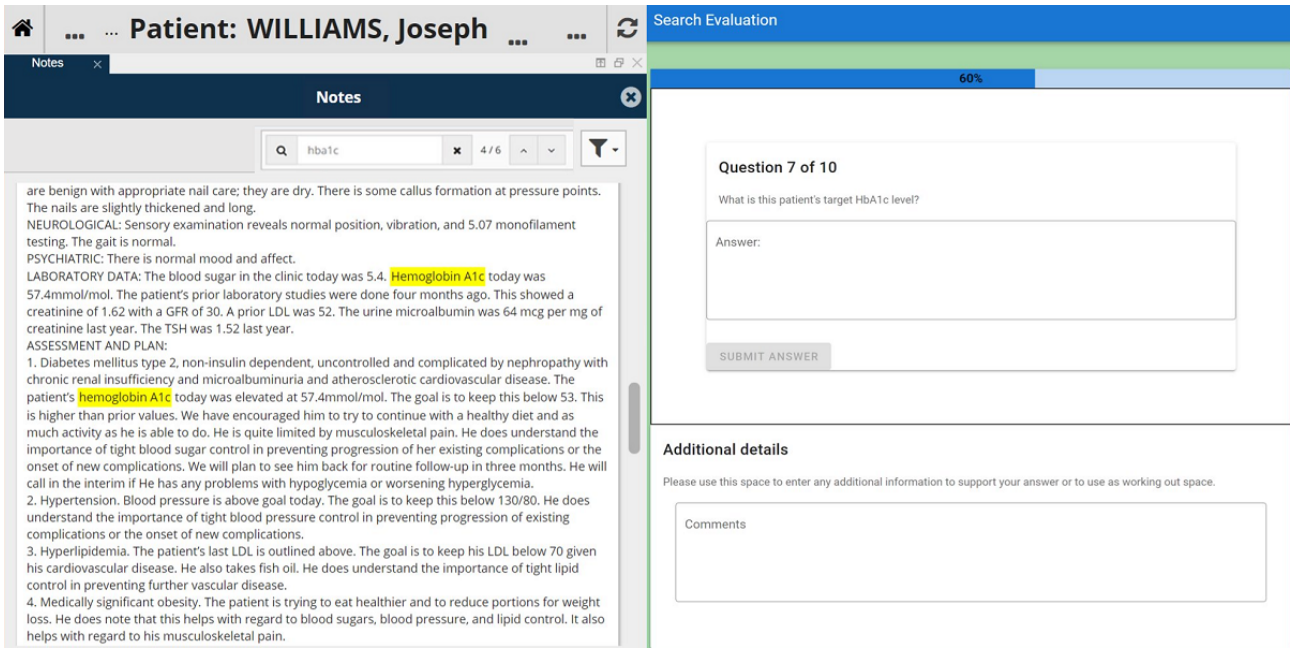
A training video was provided to evaluators, which comprised a brief introduction to the study, demonstrations of the 3 search interventions within the patient-centric viewer, and detailed instructions on how to complete the evaluation. An example patient with a small set of curated medical documents was also provided for training, on which evaluators could familiarize themselves with the capabilities of the different search functionalities.

Data Collection

Evaluators were provided with secure remote access to the evaluation environment (Figure 3), allowing the evaluation to be performed remotely from personal devices. Using this setup, evaluators could view the patient-centric viewer and the evaluation platform. Answers had to be inputted sequentially on the evaluation platform, which did not allow evaluators to return to a question once they had submitted an answer.

During each task, the evaluators submitted answers to the task questions through the evaluation platform. To ensure accurate recording of task times, evaluators were asked to perform each task in one go and to take breaks between tasks rather than during tasks. Evaluators were free to spend as long as they needed on each task. In addition, a search log was maintained, which recorded the search terms entered by the evaluator along with the search functionality used along with the time spent on each question.

Figure 3. Screenshot of the evaluation environment during a task. Evaluators only had permission to view the two relevant sites: the patient-centric viewer (left) and the evaluation platform (right). The patient-centric viewer contains the synthetic patient documents for a given patient (in this case "Joseph Williams") with "hba1c" as the search term. The evaluation platform detailed the clinical scenarios, task-specific instructions, and question-and-answer sections.



Data Analysis

Exclusion Criteria

Data were excluded where search logs showed that evaluators had used an incorrect search functionality for a given task.

Question Marking

Two clinical investigators (EP and HW) reached a consensus on the correct answers for each question. Answers were then clustered depending on the document in which they were located, and marks were awarded for finding each relevant area of correct documents. For example, if 3 pieces of clinical information across 2 unique documents were required to correctly answer a question, then 2 marks were awarded if the correct answer was inputted as the evaluator had successfully found both documents. Questions were weighted equally.

Statistical Analysis

Data analysis was performed using custom Python code. For all metrics, samples were weighted to compensate for imbalances in group size (see *Evaluator Demographics and Group Stratification*). Paired 2-tailed *t* tests were performed to determine if there was a significant difference in timing and accuracy between (1) string search and no search, (2) NLP-enhanced search and no search, and (3) NLP-enhanced search and string search. A significance level of P=.10 was applied.

Search Term Analysis

Following the study, search term logs were analyzed to extract the number and pattern of search terms for each type of search.

User Perceptions

User perceptions were assessed via a feedback survey (see Multimedia Appendix 1) which included a mix of Likert scale ratings, from "strongly disagree" to "strongly agree," and free-text responses. We clustered free-text responses by topic; we have summarized our overall findings in the *User Perceptions of NLP-Enhanced Search* section as they relate to 4 underlying questions of interest: "How was NLP-enhanced search perceived?"; "Is NLP-enhanced search better than string search?"; "Would NLP-enhanced search make clinical workflows more efficient?"; and "Would NLP-enhanced search reduce cognitive load?"

Results

Evaluator Demographics and Group Stratification

In total, 60 evaluators were recruited with multiple levels of clinical experience from medical students to doctors and from 9 specialties ranging from vascular surgery to general practice. Of 60 recruited evaluators, 44 completed the tasks; 35 were included in the final analysis (Table 2), while 9 were excluded. Evaluators were excluded from the quantitative analysis if their data were corrupted (n=2) or they completed the tasks incorrectly (n=7); for example, by using the wrong search functionality for a given task. From the original 20 evaluators per group, we observed 7 (group 1), 13 (group 2), and 15 (group 3) successful completions. There were 42 responses to the feedback survey. Table 2 shows the final distribution of clinical experience across the groups.

Table 2. Summary of allocation across clinical bands and study groups.

Clinical band	Group 1	Group 2	Group 3	Total
Medical students, n				
Preclinical (years 1-3)	4	3	3	10
Clinical (years 4-6)	0	4	2	6
Doctors, n				
1-5 years of clinical experience	0	3	3	6
6-10 years of clinical experience	1	1	4	6
11+ years of clinical experience	2	2	3	7
Total, n	7	13	15	35

Effect of Search Functionality on the Speed and Accuracy of Task Completion

The results are shown in Tables 3 and 4. Overall, NLP-enhanced search facilitated significantly more accurate task completion than both string search (5.14%) and no search (5.13%). In terms of speed, NLP-enhanced search and string search facilitated significantly faster task completion than no search (11.5% and 16.0%, respectively); there was no significant time difference between string search and NLP-enhanced search.

Table 3. Accuracy and time for different search functionalities, showing mean (SD) values across evaluators.

Search functionality	Accuracy (%), mean (SD)	Time per task (minutes), mean (SD)
None	83.8 (9.94)	20.2 (10.8)
String	83.7 (10.8)	17.0 (5.9)[a]
Natural language processing–enhanced	88.1 (9.07)[a]	17.9 (7.20)

[a]Best outcomes.

Table 4. Pairwise comparisons among different search functionalities, showing mean (SD) values in the difference across evaluators.

Search functionality comparison pairs	Accuracy increase (%)		Time difference (minutes)	
	Difference, mean (%; SD)	*P* value	Difference, mean (%; SD)	*P* value
None vs string	–0.01 (0.01; 14.5)	.93	–3.22 (–16.0; 9.78)	.006
None vs NLP-enhanced	4.30 (5.13; 13.1)	.08	–2.32 (–11.5; 7.64)	.008
String vs NLP-enhanced	4.30 (5.14; 10.5)	.02	0.91 (5.34; 5.05)	.18

Analysis of Search Terms Used by Evaluators

Analysis of the logged search terms (Table 5) revealed that evaluators tried almost twice as many search terms when using string search compared to NLP-enhanced search, and uptake of string search was slightly lower than that of NLP-enhanced search; that is, the percentage of questions for which no searches were performed was higher for string search.

The higher number of search terms required for string search might intuitively be explained by the user needing to attempt multiple synonyms to find relevant information. For instance, for the question, “Does the patient have a history of stroke?” in the text, there were 4 negative mentions scattered through the documents: “does not look like she has a stroke,” “No TIA or CVA” (ie, no transient ischemic attack or cerebrovascular accident), “No CVA,” and “No CVA.” NLP-enhanced search found all mentions with the search term “stroke” (which was the only term that evaluators attempted), but string search evaluators also attempted “TIA,” “CVA,” “neurological,” “history,” and “infarction” in their efforts to find all relevant information. Interestingly, we see that evaluators were sometimes searching for neighboring words (“history” or “neurological”) most likely as a method to bypass the possible variation in textual mentions. Further, string search does not match spelling variants (or misspellings); therefore, evaluators sometimes tried different spellings; for example, for the question, “Is the patient currently on full-dose anticoagulant treatment?” both “anti-coagulant” and “anticoagulant” were used as successive search terms by evaluators using string search.

This analysis also highlighted that the strict parameter settings for string search meant that search terms matched only to whole words, not to substrings; thus, evaluators could not search with a prefix. We observed some evidence of evaluators adjusting to this—for example, searching first for “anticoag” and then “anticoagulant” or searching for both “smoke” and “smoker”—and this also increases the number of search terms attempted.

Table 5. Analysis of used search terms showing the percentage of answers that used search and the mean (SD) values of the number of search terms for each of these answers.

Search functionality	Answers using the search functionality, %	Search terms per answer, mean (SD)
String	83.7	3.51 (2.91)
Natural language processing–enhanced	95.1[a]	2.05 (1.49)[a]

[a]Best outcomes.

User Perceptions of NLP-Enhanced Search

We used the survey shown in Multimedia Appendix 1 to gather information about user perceptions of NLP-enhanced search. Below we summarize responses under 4 headings.

How Was NLP-Enhanced Search Perceived?

Most respondents positively described the capabilities of NLP-enhanced search, noting its identification of misspellings, word forms, and synonyms, though some reported that NLP-enhanced search returned too many findings ("[NLP-enhanced] search was very clever and thorough but could return 100 results"). However, when rating the efficacy of NLP-enhanced search, 76% of respondents thought that any unrelated findings—that is, false positives—did not significantly impact the usefulness of the search algorithm.

Is NLP-Enhanced Search Better Than String Search?

Overall, 81% of respondents agreed that NLP-enhanced search facilitated more relevant IR than string search. However, many commented that the string search capabilities within the patient-centric viewer were more limited than they were accustomed to on everyday devices, stating that "string search was too discriminatory" (the parameter settings meant that only whole word matches were returned, not substrings, as discussed in the *Analysis of Search Terms Used by Evaluators* section).

Would NLP-Enhanced Search Make Clinical Workflows More Efficient?

Overall, 93% (39/42) respondents agreed that NLP-enhanced search would make clinical workflows more efficient than string search, in particular during clinics and clerking of patients. Free-text feedback reflected this, with respondents reporting that NLP-enhanced search was useful and less time consuming than string search or no search when retrieving specific information. One evaluator commented, "the [NLP-enhanced] search tool made it significantly easier for me to find the information I was looking for and also quicker." On the other hand, respondents further reported that NLP-enhanced search would not always be the best method for situations where a comprehensive overview of a patient is needed. In this case, assimilating information using manual review (no search) would be more effective. One evaluator said, "I felt that using the [NLP-enhanced] search tool meant I wasn't focussing on the case as much but just looking for words." A common opinion was that NLP-enhanced search would be a useful addition to manual review for clinical tasks.

Would NLP-Enhanced Search Reduce the Cognitive Load?

Respondents frequently mentioned that NLP-enhanced search made it easier to retrieve the information they were looking for, with one evaluator stating that "[NLP-enhanced] search is an excellent tool for a quick way to filter through relevant information." While a few mentioned that too many results were returned, respondents also reported that going through the relevant findings was easier and preferable to a full manual review of the notes, with manual review being described as "tedious," "painstaking," and "very easy to miss vital information." One evaluator commented that NLP-enhanced search could "improve the workload of an already overworked profession."

Discussion

Principal Findings

Our results showed a significant increase in accuracy when NLP-enhanced search was used compared to when string search and no search were used, while both NLP-enhanced search and string search offered time savings. There was a perception of easier navigation from evaluators and a measured decrease in required interactivity in the case of NLP-enhanced search (lower number of search terms than those obtained with string search). We caveat this conclusion with the observation that the strict parameter settings of string search meant that search terms matched only with whole words, not substrings; this increased the number of terms that evaluators used and potentially reduced the search accuracy, compared to a string search version that matches also to substrings.

There is limited literature on the potential impact of EHR search tools on day-to-day clinical care [30]. Our results support those of previous studies [20-22], which have reported that semantic search tools allow faster and more accurate EHR task completion in simulated clinical workflows. A related study [31] reported that artificial intelligence–optimized patient records improve speed in answering clinical questions while maintaining the same accuracy. Interestingly, the impact of the patient record search engine MorphoSaurus has been measured in a real-world clinical setting [32], albeit with user surveys only. This method would have had the benefit of involving real-world stresses such as task interruptions and time pressure, as well as the key element of patient interaction. Importantly, however, our method of using a controlled simulated clinical environment enabled us to control for variables such as distractions or interruptions, as well as variation in the complexity and length of medical records. Additionally, our crossover design controlled for individual participants' ability, experience, and diligence. This enabled robust comparison of quantitative and qualitative data

for each search type while minimizing the impact of confounding factors.

Overall, evaluator feedback suggested that the optimum approach to navigating clinical notes is a hybrid of manual browsing and search, depending on the context. In the real world, NLP-enhanced search is likely best employed as a complementary tool to aid clinical users in navigating clinical notes, with the ability to manually parse and ingest relevant facts from a complex medical history remaining important.

Conclusions

In conclusion, this study suggests that search tools have a positive effect on both the measured and perceived accuracy and ease of clinical IR. Search tools that can leverage NLP techniques are more effective for retrieving all relevant terms from heterogeneous medical free text. There is potential to reduce clinicians' cognitive burden and make clinical workflows more efficient. A critical direction for future research is to assess the use of search tools in real-world clinical practice.

Acknowledgments

We thank Prof Keith W Muir (Institute of Neuroscience & Psychology, University of Glasgow) for his clinical insights during the development of the natural language processing (NLP)-enhanced search tool. We would like to thank the West of Scotland Safe Haven within National Health Service (NHS) Greater Glasgow and Clyde for assistance in creating and providing a data set that was used during development of the NLP-enhanced search tool.

Many thanks to the Canon Medical Research Europe staff who developed the infrastructure required for this evaluation: Yvonne Belton, Michael Corrigan, Vismantas Dilys, Francisco Gomez, Graham Jones, Hamish MacKinnon, David Miller, Emel Muzaç, Paul Norman, and Euan Robertson. Further, we would like to acknowledge the research team that was responsible for creating the NLP-enhanced search tool: Murray Cutforth, Vismantas Dilys, Matúš Falis, Aneta Lisowska, Hamish MacKinnon, Maciej Pajak, Alison O'Neil, and Hannah Watson.

We thank our evaluators: Fiona Auld, Anna Barton, Rong Bing, Cameron Brown, Khai Syuen Chew, Jane Yi Chiam, Vanessa Chou, Luisa Ciriello, George Cooper, Iona Cutworth, Jamie Donachie, Vivienne Evans, Magdalena Gabrysiak, Eilidh Gunn, Mohamed Hamed, Hamzah Hanif, Ewen Harrison, Kylla Hernandez, Lana Huang, Katie Hunter, Haider Khan, David Kluth, Niki Kouvroukoglou, Barbora Krivankova, Tommy Le, Charles Leeson-Payne, Alinah Sum-Ping Leung, Jenny Lockhart, Jack Lugue, Angus MacLeod, Tomos Morgan, Ellen Murgitroyd, Sarah Murphy, Helen O'Neil, Yusuke Onishi, Lisa Ragunathan, Nikita Rana, Qi Shun Yong, Lucy Taylor, Evangelos Tzolos, Miriam Veenhuizen, Philippa Veenhuizen, Olivia Yu, and Sydney Zides.

We thank our pretrial evaluators: Marcus Boyd, Elizabeth Daly, Greta Economides, Keziah Lewis, Abhishek Nambiar, Sumrah Naqvi, Risako Sakatsume, Faye Sikora, and Emma Warburton.

We thank our internal Canon reviewers: William Clackett, Russell Hung, and David Miller.

We thank MTSamples for permitting free use and modification of their data to create the patient case studies.

This work is part of the Industrial Centre for Artificial intelligence (AI) Research in digital Diagnostics, which is funded by Innovate UK on behalf of UK Research and Innovation (project 104690). FV Mehendale's research at the University of Edinburgh is supported by the Caledonian Heritable Foundation.

Authors' Contributions

EHP co-designed the study, co-designed the patient histories, reviewed the synthetic patient notes, designed the tasks, designed the clinical feedback survey, organized evaluator recruitment, recorded training materials for evaluators, performed preliminary analysis of the findings, and contributed to the manuscript draft. HIW co-designed the study, co-designed the patient histories, created the synthetic patient notes, reviewed the tasks, reviewed the clinical feedback survey, supported evaluator recruitment, organized the infrastructure for the practical evaluation, contributed to and reviewed the analysis, and contributed to and reviewed the paper draft. FVM co-designed the study, reviewed the patient histories, reviewed the synthetic patient notes, reviewed the tasks, reviewed the clinical feedback survey, reviewed the analysis, and contributed to and reviewed the paper draft. AQO co-designed the study, organized provision of the NLP-enhanced search, reviewed the tasks, reviewed the clinical feedback survey, contributed to and reviewed the analysis, and contributed to and reviewed the manuscript draft.

Conflicts of Interest

HIW and AQO are employees of Canon Medical Research Europe, who provided the software and algorithms for this evaluation. EHP was sponsored by Canon Medical Research Europe during her Spring 2021 BSc research project at the University of Edinburgh ("Evaluation of a natural language processing algorithm for searching medical documents") which was the basis for this evaluation. EHP had previously performed paid annotation work for the development of the NLP-enhanced search tool.

Multimedia Appendix 1

Feedback survey which the evaluators were requested to fill out on completion of the clinical tasks.

[PDF File (Adobe PDF File), 198 KB-Multimedia Appendix 1]

References

1. Holanda A, do Carmo E Sá HL, Vieira A, Catrib AMF. Use and satisfaction with electronic health record by primary care physicians in a health district in Brazil. J Med Syst 2012 Oct;36(5):3141-3149 [FREE Full text] [doi: 10.1007/s10916-011-9801-3] [Medline: 22072279]
2. King J, Patel V, Jamoom E, Furukawa MF. Clinical benefits of electronic health record use: national findings. Health Serv Res 2014 Feb;49(1 Pt 2):392-404 [FREE Full text] [doi: 10.1111/1475-6773.12135] [Medline: 24359580]
3. Burke H, Sessums L, Hoang A, Becher DA, Fontelo P, Liu F, et al. Electronic health records improve clinical note quality. J Am Med Inform Assoc 2015 Jan 1;22(1):199-205 [FREE Full text] [doi: 10.1136/amiajnl-2014-002726] [Medline: 25342178]
4. Entzeridou E, Markopoulou E, Mollaki V. Public and physician's expectations and ethical concerns about electronic health record: benefits outweigh risks except for information security. Int J Med Inform 2018 Feb;110:98-107 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.12.004] [Medline: 29331259]
5. Kroth P, Morioka-Douglas N, Veres S, Babbott S, Poplau S, Qeadan F, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019 Aug 02;2(8):e199609 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
6. Starren J, Tierney W, Williams M, Tang P, Weir C, Koppel R, et al. A retrospective look at the predictions and recommendations from the 2009 AMIA policy meeting: did we see EHR-related clinician burnout coming? J Am Med Inform Assoc 2021 Apr 23;28(5):948-954 [FREE Full text] [doi: 10.1093/jamia/ocaa320] [Medline: 33585936]
7. Yan Q, Jiang Z, Harbin Z, Tolbert PH, Davies MG. Exploring the relationship between electronic health records and provider burnout: a systematic review. J Am Med Inform Assoc 2021 Apr 23;28(5):1009-1021 [FREE Full text] [doi: 10.1093/jamia/ocab009] [Medline: 33659988]
8. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 2016 Sep 06;165(11):753 [FREE Full text] [doi: 10.7326/m16-0961]
9. Joukes E, Abu-Hanna A, Cornet R, de Keizer NF. Time spent on dedicated patient care and documentation tasks before and after the introduction of a structured and standardized electronic health record. Appl Clin Inform 2018 Jan;9(1):46-53 [FREE Full text] [doi: 10.1055/s-0037-1615747] [Medline: 29342479]
10. Beasley J, Wetterneck T, Temte J, Lapin JA, Smith P, Rivera-Rodriguez AJ, et al. Information chaos in primary care: implications for physician performance and patient safety. J Am Board Fam Med 2011;24(6):745-751 [FREE Full text] [doi: 10.3122/jabfm.2011.06.100255] [Medline: 22086819]
11. Blijleven V, Koelemeijer K, Jaspers M. Identifying and eliminating inefficiencies in information system usage: a lean perspective. Int J Med Inform 2017 Nov;107:40-47 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.08.005] [Medline: 29029690]
12. Meystre S, Savova G, Kipper-Schuler K, Hurdle J. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2018 Mar 07;17(01):128-144. [doi: 10.1055/s-0038-1638592]
13. Sinha S, McDermott F, Srinivas G, Houghton PWJ. Use of abbreviations by healthcare professionals: what is the way forward? Postgrad Med J 2011 Jul;87(1029):450-452 [FREE Full text] [doi: 10.1136/pgmj.2010.097394] [Medline: 21459778]
14. Turchin A, Chu JT, Shubina M, Einbinder JS. Identification of misspelled words without a comprehensive dictionary using prevalence analysis. AMIA Annu Symp Proc 2007 Oct 11:751-755 [FREE Full text] [Medline: 18693937]
15. Zhou L, Mahoney LM, Shakurova A, Goss F, Chang FY, Bates DW, et al. How many medication orders are entered through free-text in EHRs?--a study on hypoglycemic agents. AMIA Annu Symp Proc 2012;2012:1079-1088 [FREE Full text] [Medline: 23304384]
16. 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 [FREE Full text] [Medline: 23304398]
17. Grabenbauer L, Skinner A, Windle J. Electronic health record adoption – maybe it's not about the money. Appl Clin Inform 2017 Dec 16;02(04):460-471 [FREE Full text] [doi: 10.4338/aci-2011-05-ra-0033]
18. Yang L, Mei Q, Zheng K, Hanauer DA. Query log analysis of an electronic health record search engine. AMIA Annu Symp Proc 2011;2011:915-924 [FREE Full text] [Medline: 22195150]
19. Ruppel H, Bhardwaj A, Manickam RN, Adler-Milstein J, Flagg M, Ballesca M, et al. Assessment of electronic health record search patterns and practices by practitioners in a large integrated health care system. JAMA Netw Open 2020 Mar 02;3(3):e200512 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.0512] [Medline: 32142128]
20. Duftschmid G, Rinner C, Kohler M, Huebner-Bloder G, Saboor S, Ammenwerth E. The EHR-ARCHE project: satisfying clinical information needs in a shared electronic health record system based on IHE XDS and archetypes. Int J Med Inform 2013 Dec;82(12):1195-1207 [FREE Full text] [doi: 10.1016/j.ijmedinf.2013.08.002] [Medline: 23999002]
21. Tawfik A, Kochendorfer K, Saparova D, Al Ghenaimi S, Moore JL. Using semantic search to reduce cognitive load in an electronic health record. 2011 Presented at: 2011 IEEE 13th International Conference on e-Health Networking, Applications and Services; June 13-15, 2011; Columbia, MO. [doi: 10.1109/health.2011.6026739]

22. Hasan S, Zhu X, Liu J, Barra CM, Oliveira L, Farri O. Ontology-driven semantic search for Brazilian Portuguese clinical notes. Stud Health Technol Inform 2015;216:1022. [Medline: 26262322]
23. Hill J, Visweswaran S, Ning X, Schleyer TK. Use, impact, weaknesses, and advanced features of search functions for clinical use in electronic health records: a scoping review. Appl Clin Inform 2021 May;12(3):417-428 [FREE Full text] [doi: 10.1055/s-0041-1730033] [Medline: 34261171]
24. mark.js. URL: https://markjs.io/ [accessed 2022-09-27]
25. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
26. List of medical abbreviations. Wikipedia. URL: https://en.wikipedia.org/wiki/List_of_medical_abbreviations [accessed 2022-01-31]
27. Medical Abbreviations & Acronyms. OpenMD. URL: https://openmd.com/dictionary/medical-abbreviations [accessed 2022-02-23]
28. Medical documents. MTSamples. URL: https://www.mtsamples.com/index.asp [accessed 2022-01-31]
29. PassMedicine. URL: https://passmedicine.com/index.php [accessed 2022-01-31]
30. Natarajan K, Stein D, Jain S, Elhadad N. An analysis of clinical queries in an electronic health record search utility. Int J Med Inform 2010 Jul 1;79(7):515-522 [FREE Full text] [doi: 10.1016/j.ijmedinf.2010.03.004] [Medline: 20418155]
31. Chi EA, Chi G, Tsui CT, Jiang Y, Jarr K, Kulkarni CV, et al. Development and validation of an artificial intelligence system to optimize clinician review of patient records. JAMA Netw Open 2021 Jul 01;4(7):e2117391 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.17391] [Medline: 34297075]
32. Schulz S, Daumke P, Fischer P, Müller M, Müller ML. Evaluation of a document search engine in a clinical department system. AMIA Annu Symp Proc 2008 Nov 06:647-651 [FREE Full text] [Medline: 18999064]

Abbreviations

CVA: cerebrovascular accident
EHR: electronic health record
IR: information retrieval
NLP: natural language processing
TIA: transient ischemic attack
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 18.05.22; peer-reviewed by J Hefner; comments to author 15.07.22; revised version received 01.09.22; accepted 07.09.22; published 26.10.22

Please cite as:
Park EH, Watson HI, Mehendale FV, O'Neil AQ, Clinical Evaluators
Evaluating the Impact on Clinical Task Efficiency of a Natural Language Processing Algorithm for Searching Medical Documents: Prospective Crossover Study
JMIR Med Inform 2022;10(10):e39616
URL: https://medinform.jmir.org/2022/10/e39616
doi: 10.2196/39616
PMID: 36287591