Usability Evaluation of an Unstructured Clinical Document Query Tool for Researchers

Gretchen Hultman, MPH¹, Reed McEwan, MS, MSSE², Serguei Pakhomov, PhD³,
Elizabeth Lindemann⁴, Steven Skube, MD⁴, Genevieve B. Melton, MD, PhD¹,⁴
¹Institute for Health Informatics, ²Academic Health Center – Information Systems,
³College of Pharmacy, ⁴Department of Surgery, University of Minnesota,
Minneapolis, Minnesota

Abstract

Natural Language Processing – Patient Information Extraction for Researchers (NLP-PIER) was developed for clinical researchers for self-service Natural Language Processing (NLP) queries with clinical notes. This study was to conduct a user-centered analysis with clinical researchers to gain insight into NLP-PIER’s usability and to gain an understanding of the needs of clinical researchers when using an application for searching clinical notes. Clinical researcher participants (n=11) completed tasks using the system’s two existing search interfaces and completed a set of surveys and an exit interview. Quantitative data including time on task, task completion rate, and survey responses were collected. Interviews were analyzed qualitatively. Survey scores, time on task and task completion proportions varied widely. Qualitative analysis indicated that participants found the system to be useful and usable in specific projects. This study identified several usability challenges and our findings will guide the improvement of NLP-PIER’s interfaces.

Introduction

Optimal requirements and design considerations for a Natural Language Processing (NLP) tool intended for use by clinical researchers are not well understood. Achieving a high degree of system usability is key to creating a product that is well accepted by target users. Usability is defined as “the effectiveness, efficiency, and satisfaction with which the intended users can achieve their tasks in the intended context of product use”¹. Usability testing is employed to help ensure that an end product is usable. Usability testing is considered an effective usability methodology that has high strategic impact². It can be defined as “the process of learning about users by observing them using a product to accomplish specific goals important to them”³. Testing typically involves representative users performing realistic tasks under typical conditions³. It is a component of user-centered designed processes that is employed to create products with high usability.

Several studies have examined usability challenges associated with search engines. One study by Dudek, Mastoram and Landoni examined the importance of usability in the evaluation of general search engines⁴. Participants were asked to perform tasks using a search engine. This study found that users valued usability when selecting a search engine⁴. Kushmiruk et al. conducted an evaluation of a new automated text summarization system called Centrifuser that was designed to enable patients and families to search for information about health conditions⁵. They also evaluated three existing search engines⁵. Think-aloud data from this study revealed that users had no clear preference for one system over another but liked different aspects of different systems⁵. Trivedi et al. evaluated NLPReViz, a tool designed to allow clinical researchers to train, revise and revise NLP models for use on clinical text⁶. Clinicians were given a tour of the interface and an introduction to the tool, then were asked to build models and discuss their experience with this tool⁶. This study found that physicians were able to use the tool to build models and provided generally positive feedback⁶.

Usability Evaluation of an Unstructured Clinical Document Query Tool for Researchers (NLP-PIER) has been previously described and is a clinical notes processing platform including an NLP query and search engine for clinical and translational researchers developed at the University of Minnesota⁷. For the purpose of this study, the user-facing application components of NLP-PIER comprise the “system” being evaluated. This system (search engine + graphical user interface) is a three-tiered web application enabling users to submit queries to an application server that transforms user input into search requests against an Elasticsearch backend holding approximately 100 million indexed clinical notes from the enterprise Fairview Health system EHR. The system is secured by an authentication and authorization layer in the application server that ensures authenticated users only have access to sets of notes that are defined externally and configured in the Elasticsearch engine. This system was designed to give clinical
researchers access to NLP capabilities for searching clinical notes in an environment that is compliant for accessing protected health information (PHI). Unstructured data contains valuable information but is difficult to “unlock” for automated secondary uses. This system fills a need by making unstructured clinical text more usable for researchers. We developed the NLP-PIER system to enable “everyday” clinical researchers to search clinical notes. Similar systems have been deployed at other institutions.

NLP-PIER has two search modes: full text searching (Figure 1) and concept searching (Figure 2 and Figure 3). We anticipate the following workflow for this application: First, the user identifies a research project in which clinical notes are needed for research. Then, the user would request access to the relevant set of notes from our institution’s Clinical and Translational Science Institute (CTSI). Participants would gain access to the note set and the application after approval. Users would be granted access to the search engine that fit their needs and would switch between the two interfaces after authenticating against an identity management system. Next, NLP-PIER is used to search through the data set, likely starting with broad searches and narrowing to more useful search terms. Lastly, users export the results out of the system for further research. Often, this research may involve using the data to compile a list of patients for chart review. Figure 2 shows what search results look like when they are returned by the system. We anticipate that this system would be most useful for research projects with data that is not easily found in structured fields. For example, searching for patients with terminal ileitis or searching for patients presenting to the emergency department with abdominal pain (tasks that require access to information hidden in the unstructured part of the EHR), but not appendicitis (easily found through diagnosis codes).

Full Text Search

This search mode relies on standard information retrieval techniques for indexed content and functions similar to search engines with which computer users are familiar. The raw text of clinical notes from the EHR were indexed for keywords by Elasticsearch using the Snowball Analyzer. The full text search interface accepts text queries using the Lucene query syntax for keywords and phrases. Parenthetical grouping of search terms and logical operators (AND, OR, NOT) allow the construction of more complex queries than the default operand (AND) assumed between terms. Sorted and paged results are returned based on a standard term frequency/inverse document frequency (TF/IDF) score for the search term(s), with the highest scoring notes listed first. We choose to pursue this design because it is familiar to users who use other search applications. Search results are presented in a list with a section of text present and a link to metadata for the note in the corner (Figure 2).

Figure 1. Full text search interface – initial screen.
Search terms in the full text interface are often instances of a medical concept: either a common synonym or a single, specific term in a controlled vocabulary. Because concepts can be represented by multiple terms (strings) in the search corpus, any single search term, or even a small set of synonyms, used in a full text search will likely fail to match documents where the concept is expressed using terms not in the search input. The United Medical Language System (UMLS) meta-thesaurus solves the problem of multiple expressions for a single concept by mapping terms from disparate vocabularies to unique concepts represented by a Concept Unique Identifier, or CUI. Because multiple terms from multiple vocabularies roll up to a single concept, searching by CUI instead of specific lexical variant amounts to a form of query expansion in the full text search sense. This form of query expansion is the basis of the conceptual search interface in NLP-PIER. CUI search functionality is enabled in NLP-PIER by indexing CUIs, identified with the UIMA-based BioMedICUS NLP system through which each clinical note is processed, with the note text.

The search box in the conceptual search interface (Figure 3) gives the user the ability to search for clinical notes containing identified CUIs using a free text, auto complete input box. As the user types in text, NLP-PIER looks for free text matches drawn from the UMLS and suggests matching terms. Suggestions can be constrained by selecting a specific UMLS vocabulary and/or semantic type from the Specialist’s Lexicon. When the user selects a suggested term, NLP-PIER displays and keeps track of the CUI so it can be used in a query against the indexed CUIs. The interface supports combining CUIs using logical operators and specifying whether the concept is used in a negated context (Figure 4). Results are displayed in the same way as for the full text interface. When making design choices for this interface, it was important to account for the experience of end users. We anticipated, based on anecdotal evidence, that users would not be familiar with this type of searching. We attempted to simplify the design for the end user and borrowed elements from designs (such as web applications) that are likely to be familiar to these users.
Objectives

The purpose of this study was to evaluate the usability and user acceptance of the NLP-PIER system through direct user testing to gain insight into interface design opportunities, user acceptance, and user preferences with the platform’s two main modes: full text searching and UMLS concept searching. These findings represent a step forward toward understanding the functionality and usability needs of clinical researchers when using a self-service NLP tool for searching clinical notes.

Methods

This study was conducted at the University of Minnesota within the Academic Health Center – Information Exchange and its associated secure data shelter which serve as key components of NLP-PIER’s infrastructure enabling clinical research in a Protected Health Information (PHI)-compliant environment.

Tasks

In order to control for variability in the use of NLP-PIER we constructed an experimental protocol designed to be a set of realistic tasks relevant to clinical researchers. Tasks were designed to identify cohorts of patients that could not easily be identified using structured data. This choice was made because NLP-PIER was designed to provide distinct and added value for information only found in unstructured format. Two physicians reviewed the protocol of constructed tasks for clinical accuracy. A software developer reviewed tasks for technical accuracy. High level examples of the tasks included: performing a search in the full text interface, running a query in the concept search interface, filtering search results to find notes written by providers, and exporting the results of queries to an external program. Table 1 contains all tasks performed in each interface.

Table 1. List of Tasks.

| Mode  | Task                                      |
|-------|-------------------------------------------|
| Search| Search "chest pain"                       |
| Search| Restrict search to notes written by providers |
| Search| Remove administrative notes               |
| Search| Save the search                           |
| Search| Restrict to 2015                          |
| Search| Rerun saved query                         |
Sessions

A convenience sample of eleven clinical research faculty participated in this study. Participants were recruited by identifying clinicians who engaged in research and sending an invitation to participate via email. Eleven participants responded and agreed to participate. Data collection sessions took place in the participant’s office and lasted approximately 30-45 minutes. For all sessions, participants were seated at a laptop computer. Screen capture software (Voila!) was used to record sessions. Participants were logged into NLP-PIER and were given a tour of the interface. Each participant was given an opportunity to ask questions. Participants were also provided with a “tip sheet” of helpful hints. This was done to partially mimic “real-world” setting, where participants would have time to test out the interface and ask questions of colleagues. The first part of the usability assessment consisted of each participant completing two sets of tasks in the full text search interface. Following completion of tasks, participants completed the SUS and raw NASA-TLX surveys.

The SUS is a widely used tool that was designed to be an efficient assessment of self reported usability. It was designed in response to a need to establish a broad measure of usability that could be used to compare usability across different contexts. The SUS is “a ten item scale giving a global view of subjective assessments of usability.” The SUS score ranges from 0 to 100 with a higher score indicating a more usable system. We selected this tool because it is a standardized and widely accepted tool for measuring self-reported subjective assessment of usability.

The raw NASA-TLX is a widely used scale designed to measure subjective impressions for workload. The NASA-TLX consists of six subscales across different domains of workload: 1. Mental; 2. Physical; 3. Temporal Demands; 4. Frustration; 5. Effort; and 6. Performance. The assumption of the NASA-TLX is that the combination of these six scales represents the abstract idea of “workload.” Participants rate workload on a scale with 21 gradations for each domain. A higher score represents a higher workload. We used this tool because it is a standardized way of measuring the subjective workload a participant experiences when using an application.

Part two of the assessment was completed in the concept search interface. Similarly, participants completed two different sets of tasks using this interface. Participants then completed the SUS and NASA-TLX surveys as well as a demographic survey and a brief interview. To analyze participant opinions and feedback, we conducted interviews with participants. In post-test interviews, participants were asked the following questions:
1a) How useful do you find the NLP-PIER system?
1b) Do you have any suggestions for making it more useful?
2a) How easy do you think NLP-PIER was to use?
2b) Do you have any suggestions for making it easier?
3) Do you have any current or future projects in which you could envision using NLP-PIER?
4) Is there anything else we should know about your experience using NLP-PIER?

Analysis

To measure the different domains of usability, we analyzed our data along five main axes. In order to measure satisfaction, we analyzed SUS scores. Efficiency was measured by the time spent on task. We measured effectiveness by analyzing task completion or reason for lack of completion. We also analyzed workload through the NASA-TLX. Lastly, we compiled participant comments and feedback. To analyze time on task and task completion, we conducted content analysis on the video recordings of the sessions. For each task, we coded the start time, end time, whether the participant successfully completed the task or not, and the reason lack of completion if applicable. Tasks were coded as either competed or failed and the reason for the decision was recorded. For each task, we performed cognitive task analysis to identify the sub-tasks involved in the task. This helped us to understand the reason for a participant failing to complete a task. Two coders (RM and GH) coded two videos (18%) and discussed all discrepancies in order to establish a standard process. Questions were discussed and resolved between the two coders. One coder was a developer and the other was an informatics researcher with experience in public health. One coder (GH) coded the remaining videos. Any outstanding questions or issues were resolved between the two coders. Interviews (n=10) were recorded, transcribed and coded using QSR International’s NVIVO 11 software. One interview was excluded due to lack of completion. Two coders (EL and GH) coded all interviews (Cohen’s kappa 0.84, percent agreement 98%). Both coders were informatics researchers with experience in qualitative analysis. For each interview question, we coded the response and any associated comments made by the participant.

Results

Participants, all of whom were practicing physicians, represented a variety of different specialties including colorectal surgery, gastroenterology, hematology, and plastic surgery. Of participants who responded, 6 were male and 4 were female. Four participants reported having practiced for less than five years post-residency, three reported practicing for five to ten years, and two reported more than ten years of practice experience post-residency. All participants were medical doctors. Participants spent an average of 15 hours per week (range 7 to 30 hours) on research activities and all reported being “average” users of technology. Two-thirds had experience requesting data from our institution’s clinical data repository. Participants were not associated with the Health Informatics Department and did not have prior knowledge about NLP-PIER.

Survey Data and Content Analysis

For the full text search interface, the SUS score was 69.4 (SD 19.8), and for the concept search interface the SUS score was 66.1 (SD 32.4). The NASA-TLX score for the full text search interface was 18.8 (SD 5.7), and the NASA-TLX score for the concept search interface was 21.8 (SD 7.7) (Table 2). Content analysis revealed wide variation in time on task (1.4 sec – 85.3 sec), and task completion percentage (9%, n=1 to 100%, n=11). Because this study represented a first step toward understanding the usability of NLP-PIER, we did not have initial estimates of how long certain tasks would take and estimated that some tasks would be faster than others. To add context, we determined what tasks participants did not complete successfully and why they failed to complete the task. We determined this by performing cognitive task analysis and identifying sub-tasks of each task. Tasks were of varying difficulty, consisting of between 1 and 7 subtasks. We noted where in the task the participant deviated from the expected path. Table 3 summarizes tasks that were completed by less than 50% of participants and reasons for lack of completion. Table 4 summarizes tasks that had the highest completion percentage.
Table 2. Least completed tasks.

| Mode      | Task                          | Mean Time in Seconds (stdev) | Task completion proportion | Subtasks                                                                 | Reasons for Lack of Completion |
|-----------|-------------------------------|-----------------------------|---------------------------|-------------------------------------------------------------------------|--------------------------------|
| Full text | Restrict search to notes written in 2015 | 85.3 (57.8)                | 1/11                      | 1. Go to help menu  
2. Find date query  
3. Click add example to search box  
4. Make modifications  
5. Click search | Participants (n=10) did not go to help menu to use a pre-made query in the help menu to restrict by date. |
| Full text | Re-run saved query            | 31.4 (29.9)                 | 2/11                      | 1. Click arrow next to all notes  
2. Click on saved query | 1 participant did not save the query  
8 did not click arrow to open menu where saved queries were |
| Full text | Clear a previous search       | 6.3 (7.6)                   | 3/11                      | 1. Clear text  
2. Clear filters (alternatively, participants could refresh the webpage) | Participants (n=8) did not clear filters |
| Concept   | Run a query                   | 21.0 (32.8)                 | 5/11                      | 1. Click add to query  
2. Click search | 4 participants did not complete previous step  
2 did not click search |

Table 3. Most completed tasks.

| Mode      | Task                          | Mean time in seconds | Task completion proportion | Subtasks                                                                 | Reasons for Lack of Completion |
|-----------|-------------------------------|----------------------|---------------------------|-------------------------------------------------------------------------|--------------------------------|
| Full text | Switch number of results displayed per page | 17.0 (12.4)           | 9/11                      | 1. Go to box next to search  
2. Click down arrow  
3. Pick 20 from list  
4. Click search | 2 could not change results because they did not refresh results in previous step |
| Full text | Export search results         | 12.7 (20.8)          | 9/11                      | 1. Click export button  
2. Click button on pop-up | 2 could not export results because they failed to refresh results in previous step |
| Full text | Open exported file            | 17.5 (19.1)          | 9/11                      | 1. Doc will automatically open in excel (can perform additional steps to change options) | 2 could not open exported file because they failed to export results |
| Concept   | Search any vocabulary         | 4.7 (6.4)            | 10/11                     | 1. Use default options | 1 used a different option from the menu |
| Concept   | Select sign or symptom        | 18.7 (10.0)          | 11/11                     | 1. Use default options | N/A |
| Concept   | Use any vocabulary            | 1.4 (3.6)            | 10/10                     | 1. Use default options | One skipped task |

Qualitative Results

In interviews, all participants expressed that NLP-PIER was easy to use. One stated “very easy, I don’t have any suggestions I think it’s very straight forward.” Two stated that the full text search engine was easier to use than the
concept search interface with one stating “I thought number one (full text searching) was super simple to do the search and number 2 (concept searching) was obviously more challenging but as I said before I would welcome that challenge if it was more precise”. All participants stated that NLP-PIER would be useful with one stating “I think it could be really useful”. Two participants expressed reservations about using the system in its current state but noted that it had potential to be useful. All participants had examples of specific clinical research and quality assurance projects in which NLP-PIER would be useful. Examples included: locating patients who present to the emergency department with facial weakness and locating pediatric patients with intestinal inflammation/colitis. Despite positive feedback, users had several suggestions for improving NLP-PIER. For concept searching, participants brought up concerns about the two-step process for running a query and understanding how negation options functioned. One participant expressed concerns about not understanding how the search engine was functioning. For the full text search mode, participants had concerns about how to run advanced queries, how to search for negated terms, and how the interface would handle misspellings. For both modes, participants had concerns about the amount of text that appears from each note in the search results, were confused about the process of refreshing results, and were confused about creating filters. Additionally, participants had questions about exporting the search results and utilizing them in future research activities. They also expressed a desire to be able to search on other types of reports such as laboratory and imaging reports.

Discussion

Our usability experiment with formal user testing of NLP-PIER provides a valuable evaluation of a self-service tool for clinical researchers with NLP capabilities for EHR notes. Survey scores were similar across the two interfaces. Because SUS is judged on a scale from 0-100, it can be difficult to interpret what the score means. Bangor et al. suggested interpreting scores on a standard letter grade scale where scores below 70 indicated that the product had substantial usability issues\(^\text{21}\). SUS scores for both interfaces indicate that interfaces have marginal usability or a “D” on a letter grade scale\(^\text{21}\). This demonstrates the need to improve usability before further deploying the system. Interestingly, participants verbally indicated that they thought NLP-PIER was usable. Further work should explore this discrepancy. While we initially spoke with researchers about their interest in the system, more could have been done prior to development to understand the needs of researchers. This includes conducting a needs assessment, workflow analysis, or interviews with participants to better understand the needs of researchers before and during the process of creating this system. Large inter-participant variability was observed for both SUS and NASA-TLX scores. This underscores the difficulty in creating a tool that is useful to a wide variety of clinicians. Factors such as age, gender, and computer experience can all affect usability and tools like NLP-PIER need to be useful to researchers with a wide variety of backgrounds and experiences.

Our usability evaluation illustrated several opportunities to improve the system for clinical researchers. We brainstormed solutions to the challenges that we identified. Table 4 summarizes challenges and proposed solutions. Despite these challenges, all participants in this study thought NLP-PIER would be useful in their research and all indicated that there were current or future projects in which NLP-PIER would be useful. This aligns with other work that suggests that this type of system is useful in helping researchers harness the data in unstructured clinical notes for secondary purposes such as research and quality improvement projects\(^\text{10,11}\). Importantly, we did not evaluate the results of user queries for completeness. Further evaluation is necessary in the future. Also, future work should focus on evaluating the existing functionality and identifying additional functionality that would be useful for researchers.
Table 4. Summary of overall usability challenges and proposed solutions.

| Issue                                                                 | Proposed Resolution                                                                 |
|----------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| Lack of ability to notice when results need to be refreshed         | Clear the previous results or overlay them with a semi-transparent background on which the need to refresh is indicated |
| Wanting number of patients to be displayed at the top of search instead of number of notes | Add optional unique patient count calculation toggle                                  |
| Wanting to being able to export list of unique patients              | Will not implement at this time, this can be done with post export data manipulation |
| Overly complex filters and unclear filter titles                    | Group filters into EHR and HL7-LOINC categories and add mouse over elements exposing definitions of each data element used as a filter |
| Not clearing filters when attempting to clear a previous query      | When user submits a query without clearing or modifying previous query’s filters, alert user that filters will be re-used |
| Wanting number of notes displayed more prominently                  | Increase font size and move to be more prominent                                      |
| Formatting of displayed note text                                  | Use formatted plain text parsed from rich text input                                  |

Our study revealed several findings related to the general needs of clinical researchers when using NLP tools in research. First, our research indicated that participants are interested in these types of tools and find them to be applicable to their work. Secondly, our research revealed several usability challenges specific to the concept search interface. Our results indicated that this type of searching presented a number of challenges and seemed unfamiliar to our participants. Future work should be done to explore clinicians’ understanding of this type of system. It is likely that many researchers require additional training before feeling comfortable using this system, and may prefer full text searching when it is sufficient to meet their needs. While not part of the formal interview, several participants mentioned informally that the duty of using the tool would likely not fall to the principal investigator but to research assistants and students. Therefore, it is necessary to make tools usable for this population of users. Our study had a number of limitations. We had small a sample size and recruited participants from one user group, thus limiting the generalizability of our findings. Additionally, this study was conducted at a single health system and may not be generalizable to other institutions.

Conclusion

This study sought to employ usability testing to evaluate a self-service search engine for its usability with clinical researchers. We identified a number of usability challenges and future mechanisms to improve NLP-PIER to address the concerns of users. Additionally, we identified barriers related to the experience and familiarity around concept searching and the system’s associated interface for end users. This study also demonstrated that substantial variation exists between different users. At the broadest level, our findings illustrate the importance of incorporating user testing and feedback in the system design process.

Acknowledgements

We would like to acknowledge the Fairview Health Services Data Reporting and Analytics team UMN Academic Health Center-Information Services, the Natural Language Processing-Information Extraction Program, the Agency for Healthcare Research & Quality (#R01HS022085 (GM)), and National Institutes of Health (#R01LM011364 (GM), #R01GM102282 (SP) and #8UL1TR000114 (Blazer)).
References

1. Schumacher RM, Lowry SZ. (NISTIR 7741) NIST Guide to the Processes Approach for Improving the Usability of Electronic Health Records. NIST Interagency Internal Rep NISTIR - 7741 [Internet]. 2010 Nov 29 [cited 2017 Jun 26]; Available from: https://www.nist.gov/publications/nistir-7741-nist-guide-processes-approach-improving-usability-electronic-health-records

2. Rosenbaum S, Rohn JA, Humburg J. A Toolkit for Strategic Usability: Results from Workshops, Panels, and Surveys. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems [Internet]. New York, NY, USA: ACM; 2000. p. 337–344. (CHI '00). Available from: http://doi.acm.org/10.1145/332040.332454

3. Barnum CM. Usability testing essentials: ready, set... test! Elsevier; 2010.

4. Dudek D, Mastora A, Landoni M. Is Google the answer? A study into usability of search engines. Libr Rev. 2007 Mar 27;56:224–33.

5. Kushniruk AW, Kan M-Y, McKeown K, Klavans J, Jordan D, LaFlamme M, et al. Usability evaluation of an experimental text summarization system and three search engines: implications for the reengineering of health care interfaces. Proc AMIA Symp. 2002;420–4.

6. Trivedi G, Pham P, Chapman WW, Hwa R, Wiebe J, Hochheiser H. NLPReViz: an interactive tool for natural language processing on clinical text. J Am Med Inform Assoc. 2018 Jan 1;25(1):81–7.

7. McEwan R, Melton GB, Knoll BC, Wang Y, Hultman G, Dale JL, et al. NLP-PIER: A Scalable Natural Language Processing, Indexing, and Searching Architecture for Clinical Notes. AMIA Jt Summits Transl Sci Proceedings/AMIA Jt Summits Transl Sci. 2016 Jul 20;2016:150–9.

8. Han H, Lopp L. Writing and reading in the electronic health record: an entirely new world. Med Educ Online. 2013 Feb 5;18:1–7.

9. Brown PJ, Marquard JL, Amster B, Romoser M, Friderici J, Goff S, et al. What do physicians read (and ignore) in electronic progress notes? Appl Clin Inform. 2014;5(2):430–44.

10. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: A report of University of Michigan’s nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). J Biomed Inform. 2015;55:290–300.

11. Horvath MM, Rusincovitch SA, Brinson S, Shang HC, Evans S, Ferranti JM. Modular design, application architecture, and usage of a self-service model for enterprise data delivery: The Duke Enterprise Data Unified Content Explorer (DEDUCE). J Biomed Inform. 2014;52:231–42.

12. Elasticsearch [Internet]. [cited 2017 Aug 1]. Available from: http://www.elastic.co/products/elasticsearch

13. Unified Medical Language System (UMLS) [Internet]. [cited 2017 Aug 1]. Available from: https://www.nlm.nih.gov/research/umls

14. UIIMA [Internet]. Available from: https://uima.apache.org/. Accessed December 10, 2015.

15. BioMedICUS [Internet]. Available from: https://github.com/nlpie/biomedicus. Accessed December 10, 2015.

16. Fact Sheet SPECIALIST Lexicon [Internet]. [cited 2017 Aug 1]. Available from: https://www.nlm.nih.gov/pubs/factsheets/umlslex.html

17. Voila! [Internet]. Global Delight Technologies; 2015. Available from: http://www.globaldelight.com/voila.html

18. Brooke J. SUS-A quick and dirty usability scale. Usability Eval Ind. 1996;189(194):4–7.

19. Hart SG. NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting. Sage Publications; 2006. p. 904–8.

20. NVivo qualitative data analysis Software. QSR International Pty Ltd.; 2015.

21. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: Adding an adjective rating scale. J Usability Stud. 2009;4(3):114–123.