Journal of Biomedical Informatics 58 (2015) S189–S196

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Ease of adoption of clinical natural language processing software: An evaluation of five systems

Kai Zheng [a,b,*], V.G. Vinod Vydiswaran [k], Yang Liu [b], Yue Wang [c], Amber Stubbs [d], Özlem Uzuner [e], Anupama E. Gururaj [f], Samuel Bayer [g], John Aberdeen [g], Anna Rumshisky [h], Serguei Pakhomov [i], Hongfang Liu [j], Hua Xu [f,*]

[a] School of Public Health Department of Health Management and Policy, University of Michigan, Ann Arbor, MI, USA
[b] School of Information, University of Michigan, Ann Arbor, MI, USA
[c] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA
[d] School of Library and Information Science, Simmons College, Boston, MA, USA
[e] Department of Information Studies, University at Albany, SUNY, Albany, NY, USA
[f] The University of Texas School of Biomedical Informatics at Houston, Houston, TX, USA
[g] The MITRE Corporation, Bedford, MA, USA
[h] Department of Computer Science, University of Massachusetts, Lowell, MA, USA
[i] University of Minnesota, Minneapolis, MN, USA
[j] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[k] Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, USA

ARTICLE INFO

Article history:
Received 16 February 2015
Revised 9 June 2015
Accepted 6 July 2015
Available online 22 July 2015

Keywords:
Usability
Human–computer interaction
User-computer interface [L01.224.900.910]
Software design [L01.224.900.820]
Software validation [L01.224.900.868]
Natural language processing
[L01.224.065.580]

ABSTRACT

Objective: In recognition of potential barriers that may inhibit the widespread adoption of biomedical software, the 2014 i2b2 Challenge introduced a special track, *Track 3 – Software Usability Assessment*, in order to develop a better understanding of the adoption issues that might be associated with the state-of-the-art clinical NLP systems. This paper reports the ease of adoption assessment methods we developed for this track, and the results of evaluating five clinical NLP system submissions.

Materials and methods: A team of human evaluators performed a series of scripted adoptability test tasks with each of the participating systems. The evaluation team consisted of four "expert evaluators" with training in computer science, and eight "end user evaluators" with mixed backgrounds in medicine, nursing, pharmacy, and health informatics. We assessed how easy it is to adopt the submitted systems along the following three dimensions: *communication effectiveness* (i.e., how effective a system is in communicating its designed objectives to intended audience), *effort required to install,* and *effort required to use*. We used a formal software usability testing tool, TURF, to record the evaluators' interactions with the systems and 'think-aloud' data revealing their thought processes when installing and using the systems and when resolving unexpected issues.

Results: Overall, the ease of adoption ratings that the five systems received are unsatisfactory. Installation of some of the systems proved to be rather difficult, and some systems failed to adequately communicate their designed objectives to intended adopters. Further, the average ratings provided by the end user evaluators on *ease of use* and *ease of interpreting output* are −0.35 and −0.53, respectively, indicating that this group of users generally deemed the systems extremely difficult to work with. While the ratings provided by the expert evaluators are higher, 0.6 and 0.45, respectively, these ratings are still low indicating that they also experienced considerable struggles.

Discussion: The results of the Track 3 evaluation show that the adoptability of the five participating clinical NLP systems has a great margin for improvement. Remedy strategies suggested by the evaluators included (1) more detailed and operation system specific use instructions; (2) provision of more pertinent

* Corresponding authors at: School of Public Health & School of Information, University of Michigan, M3531 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109, USA. Tel.: +1 (734) 936 6331; fax: +1 (734) 964 4338 (K. Zheng). The University of Texas School of Biomedical Informatics at Houston, 7000 Fannin St, Suite 600, Houston, TX 77030, USA. Tel.: +1 (713) 500 3924; fax: +1 (713) 500-3929 (H. Xu).
E-mail addresses: kzheng@umich.edu (K. Zheng), hua.xu@uth.tmc.edu (H. Xu).

http://dx.doi.org/10.1016/j.jbi.2015.07.008

onscreen feedback for easier diagnosis of problems; (3) including screen walk-throughs in use instructions so users know what to expect and what might have gone wrong; (4) avoiding jargon and acronyms in materials intended for end users; and (5) packaging prerequisites required within software distributions so that prospective adopters of the software do not have to obtain each of the third-party components on their own.

1. Introduction

Over the past two decades, the advent of new high-throughput technologies has shifted the bottleneck in biomedical research from data production to data management and interpretation. Substantial effort has focused on developing software systems that can better manage, process, and analyze biomedical data. Moreover, biomedical software also plays a critical role in improving productivity and reproducibility of biomedical studies [1]. While some recent attention has been directed toward the challenges related to locating, re-using, and properly citing biomedical software (cf. http://softwarediscoveryindex.org/report/), another important aspect is how easy it is for prospective users and user organizations to adopt these biomedical software systems. In clinical environments, the skepticism surrounding the value and cost effectiveness of health IT had been a key factor accounting for the low adoption rate of electronic health records (EHR) in the U.S. which led to significant government interventions [2,3]. Among the deployed health IT systems, the lack of usability has further hindered their effective use and contributed to numerous unintended adverse consequences such as user frustration and distrust, disrupted workflow, decreased efficiency, and escalated risks to patient safety [4–6]. However, few studies have been conducted to formally investigate the ease of adoption of software that supports biomedical research.

Recently, large EHR databases have become an enabling resource for clinical and translational research [7,8]. One challenge of the secondary use of EHR data is that much of detailed patient information is embedded in narrative clinical documents. Therefore, natural language processing (NLP) technologies, which can extract structured information from free text, have received great attention in the medical domain. Many clinical NLP systems have now been developed and widely used to facilitate various types of EHR-based studies, such as pharmacovigilance, genomic, and pharmacogenomic research [9–13]. While the target users of clinical NLP systems are often more technologically versed, they are by no means immune to poor software adoptability and usability issues [14]. Further, the lack of adoptability could limit the use of NLP systems to a small number of experts, severely undermining their potential for widespread diffusion to broader user bases.

To develop a better understanding of why there has been a lack of adoption of medical NLP tools beyond the community that develops them, a special track, *Track 3 – Software Usability Assessment*, was introduced in the 2014 i2b2 Challenge. The goal of this track was to conduct thorough adoptability evaluations – from software discovery to software installation and use – to assess how well the participating NLP systems might be received by prospective adopters. In this paper, we report the ease of adoption assessment methods that we developed for this track, as well as the results from evaluating five NLP system submissions.

It should be noted that the objective of *Track 3 – Software Usability Assessment* of the 2014 i2b2 Challenge was not to rank the participating systems based on their ease of adoption ratings. First, these systems all serve distinctive purposes and some of them, by nature, are more complicated to adopt than others. Second, the design philosophy of these systems may vary substantially according to their intended use scenarios and method of deployment. For example, some systems may choose to only provide command-line interaction modality so they can be readily invoked from other software programs; whereas some other systems provide rich graphical user interface (GUI) interfaces intended for direct interaction with end users. Thus, the results of the Track 3 evaluation should be interpreted within its own context: a higher ease of adoption rating does not necessarily suggest that a system has superior adoptability relative to the other systems evaluated.

2. Materials and methods

2.1. Scope of evaluation and submission requirements

All current and prior i2b2 Challenge participants who had developed their systems leveraging any of the i2b2 datasets since 2006 were invited to submit their work. Participating teams were only required to provide the name of the system, the URL where its descriptions and user manuals could be found, and the URL from which its executable or source code could be downloaded.

The goal of this track was to evaluate software adoptability from end users' perspective. Therefore, we only accepted systems that had a user interface (command-line or GUI); programmable components that could not be directly operated by end users, such as classes, libraries, and controls, were not included. Further, certain NLP systems offer both an online version where users may enter text or upload input files to be processed, and a downloadable version that can be locally compiled or installed. In such cases, we always chose the downloadable version to evaluate, based on the premise that a local implementation would be the preferred method for most adopting organizations due to HIPAA concerns.

2.2. Evaluators and evaluation environment

A total of twelve evaluators assisted in the Track 3 evaluation. Each of them performed a series of scripted adoptability test tasks with each of the clinical NLP systems submitted.

The two co-chairs of the track (KZ and HX) first created a draft protocol consisting of the test tasks and an evaluation instrument for collecting evaluator feedback (detailed in the next section). Two co-authors of the paper (VV and YL) then did a test run of installing and using each system. Their experience informed the further refinement of the evaluation protocol.

Their experience also led to the recognition that installing some of the participating clinical NLP systems could be a very demanding task well beyond the capability of most average users. Therefore, only four "expert evaluators," all of whom have an undergraduate or graduate degree in computer science, were asked to perform all evaluation tasks including software installation. The remaining eight individuals represent the "end user evaluators" class in the evaluation. They were only asked to work with the systems that had been preinstalled for them.

All of these end user evaluators were graduate students enrolled in the University of Michigan's Master of Health Informatics Program (http://healthinformatics.umich.edu). Six of

them have clinical degrees (two MDs, two nurses, and two pharmacists); the other two have general technologist backgrounds (e.g., business IT). Aside from being a convenience sample, this group of students was also purposefully chosen because many of them had a career projection of working in the IT department of a healthcare organization or in health IT consulting firms. These students thus approximate members on a decision-making team that makes health IT acquisition recommendations. If they have difficulties in appreciating and using the participating NLP systems, it will cast a shadow on the likelihood of these systems being widely adopted.

The evaluation environment was prepared using two Hewlett-Packard ProBook 6470b laptops with dual-core Intel i5-3360M processors clocked at 2.6 GHz. Because Linux is the preferred target platform for most of the clinical NLP systems submitted, we installed Ubuntu 14.04.1 LTS on both laptops as a virtual machine via Oracle VM Virtualbox.

We also installed a formal software usability testing tool, Turf (Task, User, Representation, and Function, http://sbmi.uth.edu/nccd/turf/), to record the evaluators' interactions with the NLP systems and their hosting websites (e.g., mouse clicks, cursor movements, and keyboard strokes). Because Turf also allows for audio recording, we asked the evaluators to 'think aloud' while performing the evaluation tasks, especially when they ran into difficulties.

Each of the expert evaluators was given 48 h to complete the evaluation tasks, typically over a weekend. They were instructed to use a clean copy of the virtual machine to install each system, to eliminate potential software conflicts and to avoid situations in which the prerequisites required for a system were already installed with another system. For the end user evaluators, we scheduled two-hour sessions with each of them. They were however allowed to use as much additional time as needed if their schedule permitted. Also, the order in which each system was evaluated was randomized. All evaluators volunteered their time for this study. The Intuitional Review Board approval was not sought because the study did not involve any human subjects. All evaluations were conducted in October 2014.

2.3. Evaluation Tasks and Evaluation Instrument

We evaluated the adoptability of each of the submitted systems along the following three dimensions: *communication effectiveness*, *effort required to install*, and *effort required to use*. These dimensions were informed by well-established technology acceptance theories which postulate that people's decision to accept (or reject) a technology was principally formed based on two perceptions: perceived usefulness and perceived ease of use [15,16].

Communication effectiveness measures how well a system communicates its designed objectives to its intended audience. It is an important factor influencing the decision-making process of prospective adopters: obviously, if a system fails to convey to its intended audience what its designed objectives are (perceived usefulness), it will unlikely be widely adopted. To assess this measure, we first asked the evaluators to find out what each system is designed to do, and report how easy it was to locate this information, and how effective this information was in helping them understand the system's designed objectives.

Further, in consumer behavior research and the innovation diffusion literature, it has been well demonstrated that consumption experience with a product or service (i.e., trialability) constitutes an important basis for purchase or adoption decisions [17,18]. Some participating systems indeed provide a trial/demo version which allows prospective adopters to see the system in action without going through potentially cumbersome steps to download, install, and configure it. We deemed this a valuable feature for enhancing communication effectiveness. We therefore asked our evaluators to report if a system provided a trial/demo version on their website, and whether it helped them understand the objectives and features of the system.

Next, we evaluated the amount of effort it requires to install a system, including the effort to install the prerequisites that must be in place for a system to run properly and to perform basic processing tasks. This is an important dimension to include because most medical NLP systems that we evaluated need to be installed locally before prospective adopters can try out the software. Note that prerequisites that can be commonly found in everyday computing environments, such as Java Runtime Environment (JRE) and Python, were preinstalled and were not counted toward the installation effort. As described earlier, the installation task was only performed by the expert evaluators. At the end of the installation session, they were asked to report how easy it was to locate the installation guide for the system, and how easy it was to follow the guide to install the prerequisites and then the system itself.

Lastly, we asked the evaluators to use each system to process a few sample medical documents. They were then asked to report how easy it was to locate use instructions, and how easy it was to process the documents and interpret the output produced.

These three adoptability dimensions were assessed through 11 questions organized under three evaluation tasks – Task 1: Evaluation of the Website Hosting the System (Questions 1–4), Task 2: Installation (Questions 5–7), and Task 3: Use (Questions 8–10). Unless otherwise specified, most of these questions used a five-level response scale as follows:

- *Effortless or nearly effortless* (2)
- *Somewhat easy but there are challenges* (1)
- *Somewhat difficult* (0)
- *Extremely difficult, nearly impossible* (−1)
- *Could not figure it out* (operationalized as "*I was not able to locate it*" or "*I was not able to get it to work*" depending on the context, −1)

Numbers in the parentheses indicate the score assigned to the system under evaluation. Note that through observing some of the evaluation sessions, we recognized that even when an evaluator decided to give up a task after repeated trials, she or he might be able to get it to work if provided with unlimited time. The last question ("*Could not figure it out*") is thus conceptually similar to "*Extremely difficult, nearly impossible*." We therefore gave the system the same score (−1) when either of these responses was selected.

At the end of instrument, we also provided an open-ended question asking the evaluators to describe their general impression of the system, or any improvement suggestions they might have, in a free-text narrative format.

Table 1 summarizes the tasks and questions included in the evaluation instrument. The full evaluation protocol is provided in Appendix A (expert evaluator copy) and B (end user evaluator copy).

2.4. Data analysis

Statistical analysis of the quantitative responses to the evaluation questions was performed using *R* version 3.1.2. A human coder analyzed the screen streams recorded in Turf first to extract the starting and ending time of each installation session. Then, the coder did a focused analysis on the 'pauses' where the evaluators appeared to have difficulties in installing or using the software, and used voice recordings to understand what the issues might be and whether/how the evaluator eventually resolved them. We also performed a qualitative analysis of the narrative feedback that the evaluators provided via the open-ended questions (Q1 and Q11).

Table 1
The evaluation instrument.

Task group	Evaluation question	Response type/scale[†]
Task 1: Evaluation of the website hosting the system	Q.1 Based on the information provided on the website, are you able to find out the designed objectives of the system? You may write below what you learned, or you may write, "I couldn't figure it out!"	Open-ended[§]
	Q.2 Is it easy to locate this information (i.e., the designed objectives of the system)?	
	Q.3 Does the website provide an online demo of the system? If so, is it easy to find the demo?	
	Q.4 If the website provides an online demo, does the demo help you understand the objectives of the system?	Yes (2); Somewhat yes (1); Somewhat no (0); No (−1)
Task 2: Installation	Q.5 Is it easy to find the instructions on how to install the system (henceforth referred to as the "installation guide")? The installation guide could be a webpage, a document (.pdf or .doc), or a readme file in the installation directory.	
	Q.6 Is it easy to follow the installation instructions to install the prerequisites?	The last response scale, "*Could not figure it out*," was replaced with "*This system does not require prerequisites other than Java or Python*" for this question. Systems that do not require nonstandard prerequisites received a usability score of 2
	Q.7 Is it easy to follow the installation instructions to install the tool itself?	
Task 3: Use	Q.8 Is it easy to find the instructions on how to use the system (henceforth referred to as the "user manual")? The user manual could be a webpage, a document (.pdf or .doc), or a readme file in the installation directory.	
	Q.9 Is it easy to follow the instructions in the user manual to use the system to process the medical documents provided?	
	Q.10 Is it easy to interpret the results generated by the system?	
Overall impression	Q.11 Do you have any suggestions on what the authors of the system can do to make it more usable?	Open-ended

[†] The response scales are as follows unless otherwise specified: *Effortless or nearly effortless*; *Somewhat easy but there are challenges*; *Somewhat difficult*; *Extremely difficult, nearly impossible*; *Could not figure it out* (operationalized as "*I was not able to locate it*" or "*I was not able to get it to work*" depending on the context.
[§] Open-ended responses were coded as follows: if the evaluator was able to articulate the designed objectives of the system with no complaints, the system received a score of 2; if the evaluator expressed explicit concerns regarding their ability/inability to understand the objectives, the system received a score of 1, or 0, depending the severity of the issue(s) reported; if the evaluator failed to articulate the designed objectives of the system, the system received a score of −1.

3. Results

3.1. Participating systems

Eight teams submitted their systems. One team withdrew before the evaluation was conducted. Two submissions were dropped because one was not an NLP system and the other was a software library that does provide a user interface. The following five systems were eventually included in the evaluation. All of them are either open-source software or are freely available under academic licenses:

- **BioMEDICUS** (The BioMedical Information Collection and Understanding System) [19].
- **CliNER** (The Clinical Named Entity Recognition System) [20].
- **MedEx** (Medication Information Extraction System) [21].
- **MedXN** (Medication Extraction and Normalization) [22].
- **MIST** (The MITRE Identification Scrubber Toolkit) [23].

Table 2 provides more detail about each of these systems.

Table 2
Participating NLP systems.

Name	Purpose	Interaction modality	Prerequisites[†]	URL
BioMEDICUS	Processing and analyzing text of biomedical and clinical reports	Command-line	Java 8 git maven uima LVG LexAccess MetaMap	https://bitbucket.org/nlpie/biomedicus/
CliNER	Named entity extraction	Command-line	python-pip (numpy, scipy, scikit-learn) python-virtualenv python-dev g++ gfortran libopenblas-dev liblapack-dev	http://text-machine.cs.uml.edu/cliner
MedEx	Medication extraction	Command-line	Java 7	https://code.google.com/p/medex-uima/
MedXN	Medication extraction	GUI	Java 7	http://ohnlp.org/index.php/MedXN/
MIST	De-identification	GUI	_sqlite3	http://mist-deid.sourceforge.net/

[†] As of October 2014 when the Track 3 usability evaluation was conducted.

3.2. Ease of adoption ratings

The quantitative ease of adoption ratings provided by the evaluators on each of the three dimensions are reported in Tables 3–5, respectively. As described earlier, the numeric scores range from −1 to 2; higher scores indicate that the system might be easier to adopt.

Table 3 shows the results on communication effectiveness. Not all participating NLP systems did a very good job conveying their designed objectives to prospective adopters. Most of evaluator complaints concentrated on the lack of specificity in the objective statement. For example, one system described its purpose as "*to provide new analytic tools for processing and analyzing text*." Several evaluators commented that this information was too general to help them get a good grasp of what the system was designed to do, and how it differed from other NLP software offerings: "*I am not sure what it says it is!*" Another evaluator further speculated that the vague objective statement of some of these systems "*does not really motivate the first-time users/speculators. It would be good to include links to related research work*."

Several evaluators also commented that the hosting website of some of the systems was too technical and was not intended for people without an extensive background in medical NLP: "*It is mainly a project maintenance website, all about the technical details of tool installation and source code. There is only one short paragraph vaguely talking about the design objective of this tool*." They also complained about the heavy usage of acronyms on these webpages, such as "i2b2" and "UIMA," which were not adequately explained on the website and were not provided with any reference links.

Among the NLP systems that we evaluated, only BioMEDICUS provided a live online demo, and MIST provided demo-like screenshots.

Table 4 shows the evaluation results on the effort required to install the software. Downloading and installing MedEx, MedXN, and MIST were straightforward and took very little time. However, two expert evaluators failed to get BioMEDICUS to work after numerous trials, and one failed to install CliNER. Among those who successfully had these two systems installed, they spent about two hours with BioMEDICUS and one hour with CliNER.

Analysis of the screen stream and 'think-aloud' data captured by Turf showed that, for both BioMEDICUS and CliNER, the major challenge was to find and install all prerequisites that they required (listed in Table 2). Some of these prerequisites proved to be very difficult to install due to the lack of documentation, or bugs or software incompatibility issues. Some of the expert evaluators were also frustrated by the fact that some websites only provided a lengthy list of prerequisites without giving any instructions, or even hyperlinks, on where to find them, how to install them, and which version to choose.

Note that when this manuscript is written, BioMEDICUS has already significantly improved their system and their website. For example, only three prerequisites are now required, instead of seven, and direct download links are now readily available on BioMEDICUS' project website. Similarly, the project website for CliNER has been redesigned to provide a detailed overview of system objectives and output examples, as well as the technology behind it. Further, CliNER installation procedure is being updated to enable the project's core functionality to be installed as a single package, separating out only the installation of external resources required to improve system performance.

Table 5 shows the results on the effort required to use the software. Overall, the ratings are rather unsatisfactory especially among the end user evaluators. The average ratings provided by the end user evaluators on *ease of use* and *ease of interpreting output* are −0.35 and −0.53, respectively, indicating that they generally deemed these systems extremely difficult to use and understand. The *ease of use* ratings provided by the end user and the expert evaluators are highly correlated, while the *ease of interpreting output* ratings are not.

Analysis of the screen stream and 'think-aloud' data recorded via Turf further revealed several areas where the end user evaluators clearly struggled. First, many of them were unfamiliar with the Linux environment and the concept of interacting with software programs through entering commands in a terminal window. For example, when provided with the following instruction to launch a program, `Run ''[UIMA_HOME]/bin/annotationViewer.sh''`, several end users typed it verbatim without realizing that Run was not part of the command, and that `[UIMA_HOME]` was a placeholder that should be replaced with the actual application path. Most negative comments surrounding results interpretation were related to the fact that some systems produced their processing output in an XML format which was very difficult for human readers to inspect.

Even though the expert evaluators had no such technical barriers, they did not deem these systems very easy to use either. Their average ratings on *ease of use* and *ease of interpreting output* are 0.6 and 0.45, respectively. While both are higher compared to the end

Table 3
Results: Communication effectiveness.

Name	Communication effectiveness on designed objectives	Ease of locating information on objectives	Availability or ease of locating a web demo	Usefulness of the web demo (if applicable)
BioMEDICUS	1.00	0.75	1.08	1.00
CliNER	1.25	1.42	−0.33	–
MedEx	1.00	1.00	−0.83	–
MedXN	1.58	1.42	−0.92	–
MIST	2.00	1.92	0.33	0.83
Average	1.37	1.30	−0.13	0.92

Table 4
Results: Installation.

Name	Average time to install (minutes)	Ease of locating the installation guide	Ease of installing prerequisites	Ease of installing system
BioMEDICUS	112	1.25	0.75	0
CliNER	58.5	2.00	0.50	0.50
MedEx	2	1.75	1.75	2.00
MedXN	6	1.25	1.50	1.50
MIST	12.5	1.25	2.00	1.75
Average	38.2	1.50	1.30	1.15

Table 5
Results: Use.

Name	Ease of locating use instructions		Ease of use		Ease of interpreting output	
	Expert	*End user*	*Expert*	*End user*	*Expert*	*End user*
BioMEDICUS	1.00	0.13	0.50	−0.50	0.00	−0.25
CliNER	1.50	0.75	−0.25	−0.75	0.25	−0.75
MedEx	2.00	0.25	1.00	−0.38	1.25	−0.63
MedXN	1.50	0.88	1.25	0.38	0.25	−0.25
MIST	0.75	0.63	0.50	−0.50	0.50	−0.75
Average	1.35	0.53	0.60	−0.35	0.45	−0.53

user evaluators ($P < 0.05$), these ratings are still low indicating that they also had considerable struggles.

3.3. Qualitative analysis results of open-ended feedback

Qualitative analysis of the open-ended feedback revealed five salient themes. They are reported in Table 6.

Over two thirds of the evaluators, both end users and experts, expressed frustration that the instructions provided with some of the systems were not very helpful in guiding them through software installation or use. This issue is particularly pronounced with command-line based operations that many non-technically savvy evaluators were unfamiliar with. Some commands required a large number of parameters, yet the purpose and usage of some of these parameters were not explained or not well explained.

Some of the instructions provided were also found to be outdated and did not work with the current version of the software evaluated. Further, the evaluators felt that several systems failed to provide useful onscreen feedback that undermined their confidence in using the software. For example, after entering a command to instruct a system to process an input file, the system printed nothing on the screen as regards whether the task was successfully executed or not. The user had to manually look into the output folder to see if an output file was correctly generated.

As a result, an overwhelming suggestion for improvement, mentioned by almost every single evaluator participated in this study, is the provision of screen walk-throughs that could help them better understand how to use the system and, more importantly, what to expect after they perform certain actions. Lastly, many evaluators also complained about the heavy usage of jargon and acronyms in the use instructions, especially in the readme files. For example, acronyms such as "CVD" GUI, "XCAS" file, and "CPE" descriptor, bear no meaning to most evaluators.

All expert evaluators agreed that the lack of instructions on prerequisite usage had been the most significant barrier to installing and using some of the systems. They could not imagine how an inexperienced user would possibly be able to figure out some very vague instructions, such as "*Start MetaMap server*," without a substantial investment of their time and energy. As one of the expert evaluators commented, "*Indeed such tools are intended for technical people to use, but the authors should keep in mind that this is not entirely intended for hardcore computer science hackers, but for medical practitioners with far less sufficient training.*"

Table 6
Qualitative themes from analyzing open-ended feedback.

Theme	Examples
Instructions are out of date, too generic, or difficult to follow	• "*The long command line inputs are quite unwieldy to use. For the command line it would be beneficial to explain the components of the syntax.*" • "*The command provided for how to run the software is very generic, but I would prefer a more detailed example on using command to analyze a specific file.*"
Lack of useful onscreen feedback undermines user confidence	• "*Very little feedback – I almost never knew if I was doing the right thing.*" • "*Even after running commands could not easily figure out what to do with.*"
Screen walk-throughs are highly desirable	• "*A lot of information to sort through in 'overview' but demo helpful. Screenshots walking through more helpful than code just displayed on screen.*" • "*The readme is extremely hard to understand. Screenshots and videos will be helpful.*"
Use of jargon/acronyms should be avoided	• "*The documentation page can be made more informative by providing links to the definition of jargon/acronyms. For example, 'CVD' GUI, 'XCAS' File, 'CPE' Descriptor, Load 'AE'. A first-time user will get lost in these acronyms.*"
Installing and using prerequisites are very difficult	• "*Too many prerequisites required which makes the system nearly impossible to install.*" • "*For each prerequisite, it is better to list the steps online, other than providing a link. Although all prerequisites are available online, and they all have somewhat good documentation, the authors should restrict from providing information in a minimalist style.*"

4. Discussion

The results of this study suggest that some of the clinical NLP systems participating in the Track 3 evaluation of the 2014 i2b2 Challenge had significant adoptability issues. Below, we summarize several common issues that surfaced from the evaluation, and discuss potential remedy strategies that may be used to eliminate or mitigate the issues.

First and foremost, many end user evaluators struggled with systems that could only be interacted with via command-line. While these evaluators had experience using terminal or DOS to run/compile software programs in their computer programming courses (e.g. Python or Java), they grew up in a GUI-dominant computing environment and do not work with terminal programs on a daily basis. Further, some concepts unique to the Linux/Unix platform, such as exporting shell variables to allow a child process to inherit the marked variables, are very foreign to them. They also had a hard time remembering to prefix bash when starting a shell program. The use instructions of some of the participating systems, however, assumed users have solid knowledge of command-line, and thus provided very little guidance on how to prepare environment variables, start a program, and manipulate input parameters.

It should be noted that we anticipated some end user evaluators would have difficulties working with Linux, and thus provided very detailed instructions on how to maneuver in the Linux environment to perform essential tasks for the evaluation (Appendix B, page 3). We also anticipated that they would have difficulties with some systems that required certain environment variables to be set prior to running the program, which was however not well described in the use instructions (the expert evaluators learned it the hard way). We included these additional instructions in the end user copy of the evaluation protocol (Appendix B, page 3–4), or prepared the environment for them before they started the evaluation session. Despite these efforts, the command-line based operation still proved to be very challenging for many of the end user evaluators. We therefore recommend that systems providing only command-line interaction modality consider including more detailed and operation system specific instructions, without assuming prospective users are all experts of the target environment. We also recommend that such systems should consolidate their software start-up scripts, whenever possible, into one single script to ease end user operation.

Another major issue we discovered from the Track 3 evaluation is that novice users could get very anxious when a system provided little or not very useful onscreen feedback after they performed certain actions. Because they were inexperienced with the software, they would not be able to tell whether the lack of feedback was because they did something wrong, or because the system might have defects or might not have been installed properly.

Therefore almost all evaluators, both end users and experts, were very particular about having screen walkthroughs as part of the use instructions so they would be assured that (1) the software does work; and (2) they know what to expect after entering a command or performing an action.

Lastly, the expert evaluators were very frustrated by the fact that some systems required a large number of prerequisites and some of these prerequisites were even more challenging to install and use than the system itself. They were also 'outraged' by the fact that some systems only listed the names of the prerequisites required without providing any guidance on how to obtain and install them, for example, "*ensure the following packages are installed on the system*" was all that was available on one of the project websites. According to the experience of the expert evaluators, some third-party websites that hosted the prerequisites were very complex and poorly documented, and not all versions worked with the clinical NLP software evaluated. We therefore strongly encourage that designers of these clinical NLP systems package the prerequisites required within their software distributions whenever possible so as to minimize prospective adopters' effort to grab each third-part component on their own.

The Track 3 evaluation has several limitations. First, the five clinical NLP systems that we evaluated are by no means representative. Nearly all of them are research systems developed at academic or research institutions which may not be primed for widespread diffusion to end user organizations. Therefore, while the evaluation results are alarming, one might hope commercial software incorporating or re-implementing the innovations introduced by these systems might be more end user friendly. Second, while the health informatics students are good approximates of members on the decision-making team in an adopting healthcare organization, they are relatively inexperienced and due to scheduling constraints, they might not have been provided adequate time to work with each of the systems. Their experience therefore might not truly represent that of the real prospective adopters. Third, in the Track 3 evaluation, we focused on perception-based measures. While a person's perceived experience with a system matters, this perception may be biased due to individual characteristics, and this effect may be magnified in this study because of the small number of evaluators involved.

5. Conclusions

This paper reports the methods and results from *Track 3 – Software Usability Assessment*, introduced for the first time in the 2014 i2b2 Challenge, that aimed to assess the ease of adoption of the state-of-the-art clinical NLP systems. Five teams submitted their work, which was carefully examined by four expert evaluators and eight end user evaluators. The results show that the adoptability of these systems is generally unsatisfactory. Expert evaluators found it very difficult to install systems that required a considerable number of prerequisites yet did not provide much guidance on how to obtain and install them. End user evaluators struggled with systems that could only be interacted with via command-line. They also struggled with vague use instructions provided with some of the systems, and the lack of onscreen feedback. Remedy strategies suggested by the evaluators focused on improving the clarity of user instructions and usefulness of onscreen feedback, and reducing the effort for prospective adopters to install each of the prerequisites required.

Conflict of interest

There is no conflict of interest for the reported study here.

Acknowledgments

We are grateful to the following individuals who assisted in evaluating the NLP systems submitted for Track 3 – Software Usability Assessment of the 2014 i2b2 Challenge: Chandra Bondugula, Lawrence Chang, Allen Flynn, Justin Gilliam, Zhaoxian Hu, Lucy Lee, Yang Li, Tracy Jia Liu, Samantha Madden, Lu Tang, Yue Wang, and Yuting Wu. We would also like to thank Kevin Cohen for providing feedback on the usability evaluation instrument, and Drs. Jiajie Zhang and Min Zhu for providing us access to the Turf usability testing tool. The laptops used in the Track 3 evaluation were provided free of charge by the University of Michigan School of Information. This study was supported in part by National Institute of General Medical Sciences Grant 1R01GM102282, National Library Medicine Grants 2U54LM008748 and 5R13LM011411.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.07.008.

References

[1] Y. Huang, R. Gottardo, Comparability and reproducibility of biomedical data, Brief Bioinform. 14 (4) (2013) 391–401.
[2] A. Boonstra, M. Broekhuis, Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions, BMC Health Serv Res. 10 (2010) 231.
[3] D. Blumenthal, Launching HITECH, N. Engl. J. Med. 362 (5) (2010) 382–385.
[4] National Research Council. Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions. Washington, DC, USA: National Academies Press; 2009.
[5] B. Middleton, M. Bloomrosen, M.A. Dente, B. Hashmat, R. Koppel, J.M. Overhage, T.H. Payne, S.T. Rosenbloom, C. Weaver, J. Zhang, American Medical Informatics Association. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA, J. Am. Med. Inform. Assoc. 20 (e1) (2013) e2–e8.
[6] E.M. Campbell, D.F. Sittig, J.S. Ash, K.P. Guappone, R.H. Dykstra, Types of unintended consequences related to computerized provider order entry, J. Am. Med. Inform. Assoc. 13 (5) (2006) 547–556.
[7] I.S. Kohane, Using electronic health records to drive discovery in disease genomics, Nat. Rev. Genet. 12 (6) (2011) 417–428.
[8] D.M. Roden, H. Xu, J.C. Denny, R.A. Wilke, Electronic medical records as a tool in clinical pharmacology: opportunities and challenges, Clin. Pharmacol. Ther. 91 (6) (2012) 1083–1086.
[9] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research, Yearb Med. Inform. (2008) 128–144.
[10] K. Haerian, D. Varn, S. Vaidya, L. Ena, H.S. Chase, C. Friedman, Detection of pharmacovigilance-related adverse events using electronic health records and automated methods, Clin. Pharmacol. Ther. 92 (2) (2012) 228–234.
[11] O. Gottesman, H. Kuivaniemi, G. Tromp, W.A. Faucett, R. Li, T.A. Manolio, S.C. Sanderson, J. Kannry, R. Zinberg, M.A. Basford, M. Brilliant, D.J. Carey, R.L. Chisholm, C.G. Chute, J.J. Connolly, D. Crosslin, J.C. Denny, C.J. Gallego, J.L. Haines, H. Hakonarson, J. Harley, G.P. Jarvik, I. Kohane, I.J. Kullo, E.B. Larson, C. McCarty, M.D. Ritchie, D.M. Roden, M.E. Smith, E.P. Böttinger, M.S. Williams, eMERGE Network. the electronic medical records and genomics (eMERGE) network: past, present, and future, Genet. Med. 15 (10) (2013) 761–771.
[12] R.A. Wilke, H. Xu, J.C. Denny, D.M. Roden, R.M. Krauss, C.A. McCarty, R.L. Davis, T. Skaar, J. Lamba, G. Savova, The emerging role of electronic medical records in pharmacogenomics, Clin. Pharmacol. Ther. 89 (3) (2011) 379–386.
[13] Ghassemi, M., Naumann, T., Doshi-Velez, F., Brimmer, N., Joshi, R., Rumshisky, A., Szolovits, P., 2014. Unfolding physiological state: mortality modelling in intensive care units. In: Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '14), pp. 75–84.
[14] W.W. Chapman, P.M. Nadkarni, L. Hirschman, L.W. D'Avolio, G.K. Savova, Ö. Uzuner, Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions, J. Am. Med. Inform. Assoc. 18 (5) (2011) 540–543.
[15] F.D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS Quart. 13 (3) (1989) 319–340.
[16] V. Venkatesh, M.G. Morris, G.B. Davis, F.D. Davis, User acceptance of information technology: toward a unified view, MIS Quart. 27 (3) (2003) 425–478.
[17] C.A. Scott, The effects of trial and incentives on repeat purchase behavior, J. Mark. Res. 13 (3) (1976) 263–269.
[18] E.M. Rogers, Diffusion of innovations, fifth ed., Free Press, New York, 2003.

[19] Y. Liu, R. Bill, M. Fiszman, T. Rindflesch, T. Pedersen, G.B. Melton, S.V. Pakhomov, Using SemRep to label semantic relations extracted from clinical text, AMIA Annu. Symp. Proc. (2012) 587–595.
[20] W. Boag, K. Wacome, T. Naumann, A. Rumshisky, CliNER: a lightweight tool for clinical concept extraction, AMIA Joint Summits on Clinical Research Informatics, San Francisco, CA, 2015.
[21] H. Xu, S.P. Stenner, S. Doan, K.B. Johnson, L.R. Waitman, J.C. Denny, MedEx: a medication information extraction system for clinical narratives, J. Am. Med. Inform. Assoc. 17 (1) (2010) 19–24.
[22] S. Sohn, C. Clark, S.R. Halgrim, S.P. Murphy, C.G. Chute, H. Liu, MedXN: an open source medication extraction and normalization tool for clinical text, J. Am. Med. Inform. Assoc. 21 (5) (2014) 858–865.
[23] J. Aberdeen, S. Bayer, R. Yeniterzi, B. Wellner, C. Clark, D. Hanauer, B. Malin, L. Hirschman, The MITRE identification scrubber toolkit: design, training, and assessment, Int. J. Med. Inform. 79 (12) (2010) 849–859.