Journal of Biomedical Informatics 60 (2016) 286–293

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Special Communication

Data management in clinical research: Synthesizing stakeholder perspectives

Stephen B. Johnson [a,*], Frank J. Farach [b], Kevin Pelphrey [c], Leon Rozenblit [b]

[a] Division of Health Informatics, Weill Cornell Medical College, 425 East 61st Street, DV-317, New York, NY 10065, United States
[b] Prometheus Research, LLC, New Haven, CT, United States
[c] Yale University, New Haven, CT, United States

ARTICLE INFO

Article history:
Received 16 March 2015
Revised 17 February 2016
Accepted 22 February 2016
Available online 27 February 2016

Keywords:
Needs assessment
Software design
Biomedical research
Medical informatics

ABSTRACT

Objective: This study assesses data management needs in clinical research from the perspectives of researchers, software analysts and developers.
Materials and methods: This is a mixed-methods study that employs sublanguage analysis in an innovative manner to link the assessments. We performed content analysis using sublanguage theory on transcribed interviews conducted with researchers at four universities. A business analyst independently extracted potential software features from the transcriptions, which were translated into the sublanguage. This common sublanguage was then used to create survey questions for researchers, analysts and developers about the desirability and difficulty of features. Results were synthesized using the common sublanguage to compare stakeholder perceptions with the original content analysis.
Results: Individual researchers exhibited significant diversity of perspectives that did not correlate by role or site. Researchers had mixed feelings about their technologies, and sought improvements in integration, interoperability and interaction as well as engaging with study participants. Researchers and analysts agreed that data integration has higher desirability and mobile technology has lower desirability but disagreed on the desirability of data validation rules. Developers agreed that data integration and validation are the most difficult to implement.
Discussion: Researchers perceive tasks related to study execution, analysis and quality control as highly strategic, in contrast with tactical tasks related to data manipulation. Researchers have only partial technologic support for analysis and quality control, and poor support for study execution.
Conclusion: Software for data integration and validation appears critical to support clinical research, but may be expensive to implement. Features to support study workflow, collaboration and engagement have been underappreciated, but may prove to be easy successes. Software developers should consider the strategic goals of researchers with regard to the overall coordination of research projects and teams, workflow connecting data collection with analysis and processes for improving data quality.

1. Objective

This study assesses data management in clinical research from three perspectives: researchers (what they believe they want), software analysts (what they believe is needed) and developers (what is required to implement the desired features). The primary motivation for this study is that the data management needs of biomedical researchers are still poorly understood [1,2]. Clinical research is a complex undertaking and support through information technology requires a wide variety of tools [3]. Even for the data collection phase, an institution may need to provide different tools for different research projects, due to factors such as the size of studies [2].

A deeper understanding of needs is necessary to understand what gaps remain in this socio-technical ecosystem [4]. The results of this study are intended to inform the development, adoption and integration of data management software tools for clinical researchers. By attaining a clearer picture of the needs of researchers, we can begin to assess the adequacy of existing tools, identify unmet needs, and more effectively organize software tools and human processes within the ecosystem.

* Corresponding author. Tel./fax: +1 646 962 9403.
E-mail address: johnsos@med.cornell.edu (S.B. Johnson).

http://dx.doi.org/10.1016/j.jbi.2016.02.014
1532-0464/

2. Background and significance

Biomedical research is generating data at increasing volume, variety and velocity, requiring improved tools, policy and training [5]. Clinical research raises many additional data management challenges, due to the involvement of human subjects, including participant recruitment, informed consent, visit scheduling and the protection of confidentiality [3,4,6]. Data management is a crucial function in clinical research; poor management practices pose risks to researchers, participants and patients [7].

There is very little research that directly examines the needs of clinical researchers and how this knowledge would inform the development of informatics solutions [1,8]. Some related work focuses on technical aspects, such as integration of data across different systems [9], information storage and representation [10] and procedures for validation and change control [11]. Only a few studies have considered human factors, such as individual workflow [12,13], training [14], organizational incentives [15], and how users interact with software tools [16].

The discipline of software engineering has recognized that qualitative studies, potentially mixed with quantitative methods, are important for understanding the complexity that arises at the intersection of human processes and technological systems [17–20]. Qualitative and mixed approaches have provided significant insights into work practices and cultural settings in which software is embedded, but are still relatively underused [21,22]. Based on these findings, there is strong need to obtain a deeper understanding about how clinical researchers are managing their data, and what specific features of software would improve that process.

This study employs a mixed-methods approach that combines content analysis and feature extraction from interviews with surveys of three different stakeholders. The innovation of the approach is the use of sublanguage analysis to link all of the assessments together [23,24]. Sublanguage analysis has been used extensively in the structuring of clinical texts [25,26]. While sublanguage analysis has been conducted on the text of survey items [27], it has not, to our knowledge, been applied in qualitative studies for content analysis. In this study, we use a common sublanguage to represent the activities and attitudes of researchers, as well as software features from the perspective of business analysts, software developers and researchers. The common sublanguage provides a powerful mechanism for synthesizing the findings of these different assessments. These methods are described in greater detail in the following section.

3. Materials and methods

3.1. Research participants

This study investigated needs for data management in clinical research through three stakeholders: researchers who conduct clinical studies; business analysts who work with researchers to assess needs and customize software; and software developers who create and modify data management software. Business analysts were included to provide an alternate perspective on the needs of researchers. Developers were included to provide information about the difficulty involved in implementing different aspects of the software system.

Researchers were recruited at the Yale Child Study Center, Weill Cornell Medical School, University of Missouri and Boston Children's Hospital. These institutions were selected because they are large and complex, with many different research groups working with diverse data sets and tools. Each participant had to be part of a research team studying human behavior with one of the following roles: principal researcher, research scientist, postdoctoral fellow, resident, data manager, statistician, research assistant or coordinator. No restrictions were made regarding what software tools were being used by the researchers. Researchers were contacted by e-mail using snowball sampling (contacts were asked to suggest other contacts) [28]. We obtained a sample size of 22, which was is considered sufficient to attain saturation in the content analysis [29].

Because it is very challenging to identify and enroll business analysts and software developers working in clinical data management, we partnered with a software vendor (Prometheus Research, LLC) and recruited a convenience sample from their employees and business associates. We recruited 8 analysts, who had to have completed all modules of company training. We recruited 7 developers who had done contract work with the vendor, 4 of which were working from remote locations. The developers were all familiar with contemporary software standards, including Web services architecture, hypertext transfer protocol (HTTP), representational state transfer (REST) and structured query language (SQL).

3.2. Human subjects protection

Research protocols were approved by the institutional review boards at Yale and Weill Cornell Medical College. Informed consent was obtained from all participants (researchers, analysts and developers) by telephone. Recruitment information (name, telephone and email) was kept separate from data collected for analysis. Interviews were conducted over a secure phone line, transcribed by a professional service, redacted of identifying information and stored on a secure server. Recordings were destroyed after coding was completed. All surveys were conducted via a secure online survey tool.

3.3. Study design

The process for data collection, analysis and synthesis is shown in Fig. 1. Interviews with researchers were recorded and transcribed. A team of coders performed content analysis on the transcriptions using the method of sublanguage analysis. In parallel, a senior business analyst extracted potential software features from the transcriptions. The analyst's technical descriptions of features were translated into the sublanguage grammar obtained from the content analysis. This common sublanguage could then be used to translate the software features into survey items for the three different stakeholders (researchers, business analysis and software developers). Finally, the common sublanguage facilitated synthesis of the results from the three surveys with the original content analysis. Each of these steps is described in more detail in the following sections.

3.4. Researcher interviews

A research assistant conducted interviews with researchers over the telephone using a semi-structured script (Table 1). The questions were designed to get a general sense of the work performed by researchers, the perception of data management, what they felt works well, what does not work well and what is desired for the future. Recordings of the interviews were transcribed by a professional service.

3.5. Content analysis

Four coders performed content analysis on transcriptions of the interviews. Coders had a mixture of educational backgrounds (1 Bachelor's, 2 Master's and 1 Doctorate), and all had prior experience with qualitative coding. Interviews were coded sequentially, with interviews 1–4 coded by 3 coders, and interviews 5–22 by 2 coders. Inter-rater agreement was assessed using Krippendorff's