Applying a Participatory Design Approach to Define Objectives and Properties of a "Data Profiling" Tool for Electronic Health Data

Hossein Estiri, PhD[1,2], Terri Lovins, BA[1,3], Nader Afzalan, PhD[5], Kari A. Stephens, PhD[1,2,4]
[1]Institute of Translational Health Sciences, [2]Department of Biomedical Informatics & Medical Education; [3]Human Centered Design & Engineering; [4]Department of Psychiatry & Behavioral Sciences; University of Washington, Seattle, WA; [5]Department of Geographic Information Science, University of Redlands, Redlands, CA

Abstract

We applied a participatory design approach to define the objectives, characteristics, and features of a "data profiling" tool for primary care Electronic Health Data (EHD). Through three participatory design workshops, we collected input from potential tool users who had experience working with EHD. We present 15 recommended features and characteristics for the data profiling tool. From these recommendations we derived three overarching objectives and five properties for the tool. A data profiling tool, in Biomedical Informatics, is a visual, clear, usable, interactive, and smart tool that is designed to inform clinical and biomedical researchers of data utility and let them explore the data, while conveniently orienting the users to the tool's functionalities. We suggest that developing scalable data profiling tools will provide new capacities to disseminate knowledge about clinical data that will foster translational research and accelerate new discoveries.

Introduction

The increasing adoption of information technology in health care settings, supported by the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, has boosted the throughput of Electronic Health Data (EHD) over recent years[1]. The overflow of health data offers enormous potential for improved secondary use of EHD in translational research[2,3]. To pioneer new discoveries, biomedical researchers are utilizing new data tools to learn about and access EHD[4]. Advances in data science methods and technologies have resulted in the emergence of new data tool developments to translate data into actionable knowledge to address a variety of purposes.

In biomedical informatics, these new tools are gradually replacing (or complementing) current or "older" tools and are distinct from their predecessors in that they provide capabilities to process data that are larger in volume and more complex in information architecture. Tools such as cohort selection tools developed in the late 2000s to the early 2010s were not designed for the so called "Big Data." Their usage, as a result, for identifying patient cohorts from EHD is already often a prolonged and costly process[5], and can become more inefficient considering the expected overflow of EHD in Clinical Data Repositories (CDRs) in the next few years.

In addition to providing more advanced technologies, new data tools need to offer novel approaches to promote the application of EHD for research and health policy[6]. Large data sharing network efforts, such as PCORNet, MiniSentinel, eMERGE, and the NIH Collaboratory's Distributed Research Network, do not offer clear, hurdle free data profiling tools for their systems that would allow researchers to easily and quickly explore existing data networks for research application. Researchers need tools to acquire knowledge about data, before they apply other tools to analyze them. This knowledge is currently being disseminated slowly through researchers' professional networks and communication channels[4]. Further, due to governance and privacy concerns, current approaches to building EHD tools often present hurdles that prevent uptake by requiring time intensive steps (i.e., logins, training requirements). We present this approach as "data profiling." Offering asynchronous data profiling, i.e. the sharing of information about the depth and breadth of a given dataset, so users can orient to the dataset more efficiently, meaningfully, and when they have time, fulfills a critical need for researchers seeking to leverage EHD for research.

Many of the newly developed EHD interactive tools have not capitalized on user-centered design methods to meet user needs and preferences[4]. Current data tool designs are often driven by developer based assumptions about user needs and do not leverage design methodologies that have become common in the technology industry, creating a gap in design and limiting utility, usability, and uptake – e.g., features have low usability and do not address basic needs of users (i.e., addressing what, when, and where about the data before doing more in depth querying).

Participatory Design (PD) is a user-centered design approach that offers solutions to tool design in health informatics[7]. PD seeks to design *with* users as opposed to designing *for* users, actively involving both end users, as well as other stakeholders in designing tools, interfaces, or systems[8]. In health care, PD has been applied in

designing health information systems mainly as a way of addressing the multiplicity and inter-organizational nature of modern health care informatics projects[7]. For example, PD has been used: to design small scale health information systems[9]; as a knowledge generator between users, stakeholders and designers[7]; to design technically usable systems and as a research tool to perform needs assessments[10]; to design community-based health services[11] and patient-centered post-hospital transition intervention[12]; and to develop eHealth applications[13], and assistive technologies[14].

We conducted a qualitative research study, using a participatory design approach with biomedical researchers across three organizations to define concrete features or characteristics of a data profiling tool for EHD. We focused on a use case involving data derived from primary care electronic health record systems within a regional practice based research network. Based on the recommended features or characteristics, we extracted a defining set of overarching objectives and properties for the data profiling tool.

Methods

The Innovations and Collaborations Workshops

Several methods and techniques do commonly appear in the successful use of PD in the health care field. These include utilizing a workshop model in the execution of PD[7] which we used to elicit participation from a varied group of stakeholders in order to capture multiple viewpoints and experiences[9], and include tangible hands-on design activities during workshop sessions[10]. We held three participatory design meetings, which we termed "Innovations and Collaborations (IC) workshops," across the three partner institutions of the University of Washington's Institute of Translational Health Sciences (ITHS): University of Washington, Fred Hutchinson Cancer Research Center, and Seattle Children's Hospital. The IC workshops took place in parallel to a team effort that was building a data profiling tool aimed at assisting researchers with learning about and exploring an EHR data repository. Workshops were designed to play the role of an innovation incubator for the data profiling tool design effort, as a means to collect new ideas, to increase interdisciplinary collaborations, and stimulate innovations in developing the future generations of the tool.

Each workshop drew 10-14 health research experts (N = 36) with expertise in various fields related to biomedical research – including, clinical researchers, data scientists, biostatisticians, and biomedical informaticists. Existing biomedical informatics and biostatistics groups were solicited at each institution to invite participants to attend the workshops. Participation was voluntary and attendees were invited via email invitations circulated through the institutions informatics-oriented listserv. This method of invitation resulted in attracting attendees that mostly had past experience using electronic health data as either researchers, technical developers, or both.

Each workshop was 1.5-2 hours in duration, divided into three sections. During the first 30-45 minutes, attendees learned about our goal to design an EHR-driven data profiling tool for a practice based research network of federated primary care based aligned clinical data repositories, Data QUEST[15], and a prototype version of the data profiling tool that we had designed based on our expert assumptions[16], and outcomes of needs assessment studies we conducted in summer 2014[4]. Workshop attendees then formed into groups of 3-5 members and were given 30-45 minutes to collaborate and innovate. Team activities were guided by open-ended questions to collect ideas for a useful, attractive, and creative data profiling tool – examples of the open-ended questions are:

- What would be a useful/attractive/creative way for the tool to increase the users' confidence in data integrity?
- How could the tool help the users find out whether or not the data have what the users are looking for?

After the team activity, participating teams then presented their ideas, solutions, and/or their concerns in the last section of the workshop.

Content Analysis

Presentations were videotaped for post-workshop content analysis. Considering the teams' recommendations as the unit of analysis, we employed content analysis to characterize and cluster similar recommendations, categorized recommendations under derived objectives and properties, and inferred implications from recommendations' content[17,18].

Results

Over the course of three workshops, 36 attendees worked in 12 teams and made 44 recommendations. We characterized the teams' input into 15 recommended features or characteristics and identified a set of overarching objectives and properties for the data profiling tool. We first describe the overarching objectives and their underlying recommended feature or characteristics. Next we describe the overarching properties of the data profiling tool.

Our analyses led to the identification of three overarching objectives with associated features or characteristics of the data profiling tool (Figure 1). The overarching objectives of the data profiling tool should: Objective 1 - inform users of data utility, Objective 2 - enable users to explore data, and Objective 3 - easily orient users to its functions. We classified the teams' recommendations as either *features* or *characteristics*. Features were recommendations that pointed to functions that the teams suggested be built into the data profiling tool. Characteristics were recommendations that referred to design attributes of either the tool as a whole or its individual features. Some of the characteristics were followed by feature examples. Also, some of the features or characteristics had dual functionality, meaning they addressed more than one objective.

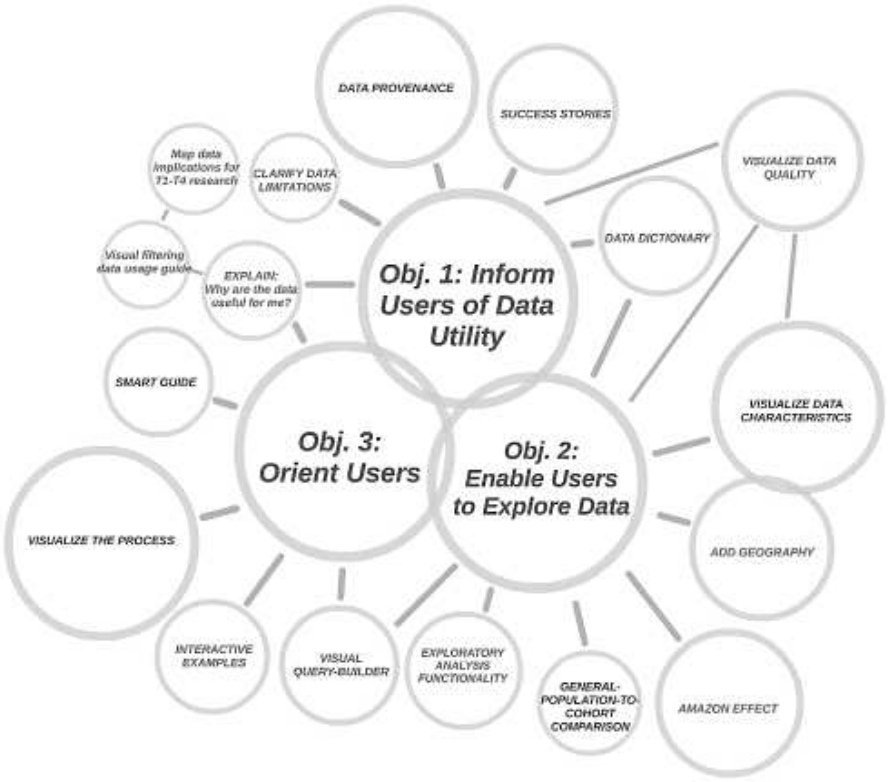

Figure 1. Overarching objectives and recommended features or characteristics of a data profiling tool. Outer circle diameter indicates the number of times each feature or characteristic was recommended, the bigger the circle, the more the feature or characteristic was recommended.

Objective 1: Inform Users of Data Utility

Teams recommended that a data profiling tool should inform users of the data utility (or lack thereof) for their respective needs, as the first requirement for the users to assess fitness-for-use. In the context of biomedical informatics, and based on our characterization of the users, a data profiling tool should inform biomedical researchers whether or not the electronic health data profiled meets their needs (has good fit-for-use) in investigating specific clinical research questions. Below are the teams' recommendations of features and/or characteristics that addressed this objective.

1.1. Explain: Why are the data useful for me?
According to the teams, an important characteristic of the data profiling tool is that it should be explicit as to why the data are useful to the researcher. Answers to the following subsidiary questions was recommended to help define this informative characteristic:

- Why do I want a cohort from your data?
- What are your data offering that I cannot easily get elsewhere?
- Are patients in your datasets accessible for my clinical trial?

Since answers to these questions can differ across researchers (by their needs, questions, research design, and status), one team recommended mapping data utility for the researchers across each of the four steps of translational science from basic science to dissemination (i.e., T1→T2→T3→T4)[19].

To answer the overarching "why" question (i.e., why are the data useful for me?), different means and formats can be utilized – i.e., from simple test description, to graphics, to tutorials. However, providing too much information can be distracting. An efficient recommendation to this concern was a feature that can provide information in steps, enabling the users to filter through and reduce information to only what is needed. This feature can be combined with a data dictionary feature and/or be used to visually demonstrate the data to the users. For instance, users can click on patients, then find out about what type of information are available about patients – for example, the user can learn that demographics, lab results, and diagnosis records are available for patients. Next, clicking on diagnosis will illustrate the type of questions that can be answered with available diagnosis data. These sets of recommendations also overlap with the third objective, orient users to the tool's functions, explained below.

1.2. Clarify Limitations
In addition to demonstrating what can be done with the data, one team suggested that it is an important characteristic for the data profiling tool to clearly communicate to the users what cannot be done with the data. The data profiling tool, therefore, needs to demonstrate both what purposes the data fit well and what purposes the data do not fit well.

1.3. Explain Data Provenance
Four out of the 12 teams emphasized the importance of clearly conveying data provenance (i.e., how the data were created and where they come from). A feature idea for this recommended characteristic was to develop a visual way to illustrate data provenance.

1.4. Success Stories
Providing users with examples of how the data and/or the tool have been used successfully in the past, or are being used currently, seemed important to 25% of the teams. A feature was proposed to include demonstration of success stories within the tool to improve informed use of both the data and the tool.

1.5. Visualize Data Quality
Workshop attendees recommended a feature to visualize different aspects of data quality, given its large relevance to utility of the data, to help users know more about the data and learn about potential issues that they might encounter using the data. Suggestions included visualizing data consistency and variability as well as establishing a mechanism that will allow users to leave reviews on data quality. Visualizing data quality also serves the second objective of the data profiling tool, to enable users to explore the data.

1.6. Data Dictionary
Participants suggested a data dictionary would be a useful feature that enhances the data profiling tool's capacity to inform users of data utility. This recommended feature also served the second objective, to enable users to explore the data.

Objective 2: Enable Uses to Explore the Data

A data profiling tool needs to enable researchers to explore depth and breadth of data with ease. Due to efficiency and privacy reasons, exploring the data, in this context, is not equivalent to letting users query data freely. Considering the privacy issues related to health data, the teams made a number of feature or characteristic recommendations that will allow clinical researchers to explore different dimensions of data. Below are the features or characteristics that the participating teams recommended across the three workshops.

2.1. Visualize Data Characteristics
Visualization was the most highly recommended characteristic of the data profiling tool for conveying information to users. Five out of the 12 teams suggested visualizing various aspects of data characteristics, so that users can more

easily learn about the data. Various forms of visualizations were recommended, including charts and maps to visualize population, and interactive visual demos. Several teams recommended visualizing the depth (date range) and breadth (inclusivity – e.g., number of clinics, population, etc.).

2.2. Cohort-population Comparison

Two teams recommended a feature that enabled users to learn about the data better by understanding how the cohort they are interested in compared with the general population that the cohort was expected to represent. For example, the feature could demonstrate that a user-selected patient cohort was 70% male compared to 48% in the general population – i.e., the cohort's gender composition was not representative of the general population.

2.3. The Amazon/Netflix Effect

A few teams at different workshops suggested a feature (that they called the "Amazon/Netflix effect") for increasing users' ability to explore the data. The idea spoke to a learning system in which users would receive suggestions about similar search topics that they might be interested in based on data collected from other users doing similar searches.

2.4. Add Geography

Teams across all three workshops recommended adding a geographic search feature (or spatially-enabled exploration characteristics) to the data profiling tool. Such a characteristic in a data profiling tool would enable users to include a spatial dimension in their data exploration that could also resonate with the "Cohort-population Comparison" and "The Amazon/Netflix Effect" recommendations. Map visualizations, showcasing patient populations across geography, were also highlighted by multiple teams as being an interesting feature.

2.5. Exploratory Analysis

Two teams suggested that learning about data will be enhanced by features that enable levels of exploratory (hypothesis-generating) analyses. Two specific examples were given. First, some correlations/relationships among selected variables could be pre-computed and used to give information about additional variables that might be correlated with the selected variable. For instance, users interested in blood pressure would receive a message about other variables that are correlated with high blood pressure within the dataset. Second, a simple or complex (e.g., Natural Language Processing) text mining feature will allow users to learn more about the data with more flexibility – e.g., receiving an aggregate count estimate of patients with particular clinical conditions from free-text data.

Objective 3: Orient Researchers

In addition to what the data have to offer to the user (clinical researcher, in this context), a data profiling tool needs to be able to clearly orient users to what features the tool has to offer. A few of the features or characteristics that the teams recommended throughout the three workshops spoke to the need for the tool to help the users understand its functionalities, critical to good usability. A data profiling tool that meets this objective would help the users learn more about the data, by learning how to make the best use of the tool. The following features or characteristics recommended by the teams fall under this third objective.

3.1. Visual Query-builder

A suggested characteristic was building data queries that could be enhanced via visual interfaces. A visual query-building functionality would improve both the user's learning experience as well as the tool's ability to orient users. In order to orient users, the visual query-builder would need to demonstrate how cohort population would change by adding new inclusion and/or exclusion criteria to define an aggregate count query.

3.2. Visualize the Process

Half of the teams across the three workshops recommended visual tutorials about the data profiling tool (i.e., how this tools works). Visual tutorials could include traditional instructional videos on how to navigate through the tool. However, most participants suggested interactivity as one of the prominent characteristics of the visualizations to further engage users with the tool. A dynamic display of how the tool was being used, an activity scroll at the bottom of the tool, or a consort diagram would let users know where they are during each step of using the tool.

3.3. Interactive Examples

Interactivity was among the most recommended characteristic for data profiling tool design. Specifically, two teams in one workshop recommended functions that provide interactive examples of how the data or the tool could be used. Interactive examples could be built using a combination of story-telling and data visualization.

3.4. Smart Guide

Another feature that was recommended by the teams was a smart guide to orient researchers to the tool, as well as the data. Two recommendations were made for building a smart guide. A few teams recommended the tool use a survey mechanism at the beginning of the process and at the exit point to learn about user needs and guide the user through the tool's features using a decision tree algorithm. This information could also be used by the development team to improve the tool and add new features. Another group recommended a functionality that would give users the ability to filter through the provided guiding information about the tool's features, based on their needs.

Overarching Properties of the Data Profiling Tool

In addition to the three objectives, from the recommendations made by the teams, we also identified five overarching properties of the "ideal" data profiling tool. Each of the 15 recommendations was associated with at least one of the overarching properties. To understand which property was of highest importance to our workshop attendees, we calculated the number of times each characteristics was indirectly suggested through participant recommendations, and weighted the raw values – divided by total number of times participants made a recommendation (44). Results are illustrated in Figure 2.

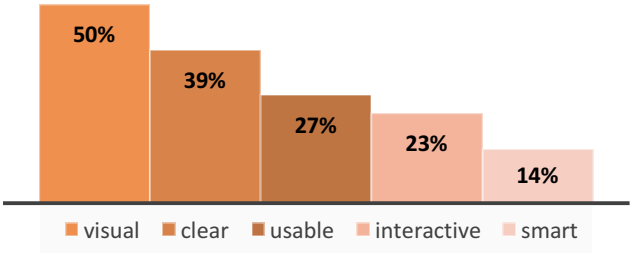

Figure 2. Ranking the five overarching properties of the data profiling tool

The data profiling tool, according to our results, should be visual, clear, usable, interactive, and smart. These properties can be applied to any of the three objectives in the design of a data profiling tool. Being visual was referred to 50% of the time a recommendation was made and related to a range of content, from visualizing the data to the process. Being clear as to what the data profiling tool offers, and what it does not offer (in terms of data, usage, and services) was the second most important property of the data profiling tool. The data profiling tool also needs to have high usability, from a user design perspective, which speaks to the importance of applying user-centered design methodologies in designing the tool. In addition to highly usable, being interactive would increase user engagement with the data profiling tool's features and functions. New technological advancements around predictive algorithms have increasingly been providing new opportunities for data tools to become smarter. Utilizing data from users, the data profiling tool needs to be smart to effectively orient users and increase its usability.

Discussion

The upturn in throughput of Electronic Health Data (EHD) from a variety of sources provides huge promise for future translational research and speeding discoveries. In order for this promise to be realized, EHD users (in this case researchers) need to gain sufficient knowledge about the available data. However, dissemination of this much-needed knowledge about data is not keeping up with the rate that electronic health data are increasing in volume and complexity.

To address this gap in dissemination of knowledge about data, our participatory design study provided guidance on features or characteristics for EHD-based data profiling tools for biomedical researchers, which we presented under three overarching objectives. Four of the recommended features and characteristics served multiple purposes and therefore overlapped between the three objectives. These overlaps refer to the synergies between the three objectives. For example, informing users of data utility can be conveyed by giving the users the ability to explore the data, and vice versa.

Workshop outcomes also led to identification of five overarching properties for the data profiling tool. Being visual was the most prominent property for the tool. Over the past few years, visualization has gained significant attention from data science audiences in general, and subsequently, bioinformatics communities – especially, with the advent

and progression of new interactive data visualization tools. However, in light of our results (e.g., that, in addition to being visual, there are also four other important properties for data profiling), we argue that while data visualization is an effective means to facilitate learning about data, at least in the context of clinical data, it may not be sufficient on its own. We characterize data visualization, in this context, under the umbrella of the broader concept of data profiling that concurrently embraces multiple properties, such as clarity and usability that, according to our results, are collectively as important as visualization when profiling clinical data. An EHD-based data profiling tool needs to be visual, but also clear, usable, interactive, and smart. These characteristics are driven from a group of EHD users that understood the specific considerations/limitations related to potentially sensitive health data and had experience using currently available tools. Therefore, some or all of these characteristics might well be generalizable to data profiling tools in other contexts.

Conclusion

While common ontologies are prevailing, like the Observational Medical Outcomes Partnership (OMOP) common data model used in our practice based research network, new scalable tools are being developed to profile different dimensions of EHD from different sources. A participatory design approach to collect recommendations for a data profiling tool to capture expert recommendations can improve the quality of data profiling tools. We compiled recommendations that ranged from simple verbatim guidelines and features like inclusion of a data dictionary and cohort-population comparator, to more complex and smart features such as smart guides and the Amazon/Netflix effect. The ideal data profiling tool can be described as a visual, clear, usable, interactive, and smart tool that is designed to inform clinical researchers of data utility and let them explore the data, while orienting the users to its functions. Future data profiling tools incorporating these recommendations should enable health researchers (the users of such tools) to smoothly evaluate the data's fitness-for-use independently, free of typical barriers to access, and without the need of human expert intervention that often bottlenecks evaluation. User testing would further strengthen the tool design by ensuring that features or characteristics of the tool create a satisfactory experience for the users.

Acknowledgements

This project was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000423. The authors would like to thank Shoonya Mohanrao, Paul Litwin from Fred Hutchinson Cancer Research Center, and Dr. Eric Tham and Daksha Ranade from Seattle Children's Research for their assistance in holding the workshops, and all workshop participants who gave us invaluable ideas. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Arts DGT, De Keizer NF, Scheffer G-J. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc*. 2002;9:600-611. doi:10.1197/jamia.M1087.
2. Hersh WR. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. *Am J Manag Care*. 2007;13:277-278.
3. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20:144-151. doi:10.1136/amiajnl-2011-000681.
4. Stephens KA, Lee ES, Estiri H, Jung H. Examining Researcher Needs and Barriers for using Electronic Health Data for Translational Research. *AMIA Summits Transl Sci Proc*. 2015;2015:168-172. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525243/.
5. Fernández-Breis JT, Maldonado JA, Marcos M, et al. Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. *J Am Med Inform Assoc*. 2013;20(e2):e288-e296. doi:10.1136/amiajnl-2013-001923.
6. National Research Council. *The Learning Healthcare System: Workshop Summary (IOM Roundtable on Evidence-Based Medicine)*. Washingotn, DC; 2007.

7. Scandurra I, Hägglund M, Koch S. From user needs to system specifications: Multi-disciplinary thematic seminars as a collaborative design method for development of health information systems. *J Biomed Inform*. 2008;41(4):557-569. doi:10.1016/j.jbi.2008.01.012.

8. Sanders EB. From user-centered to participatory design approaches. In: *Design and the Social Sciences: Making Connections*.; 2002:1-8. doi:10.1201/9780203301302.ch1.

9. Revere D, Dixon BE, Hills R, Williams JL, Grannis SJ. Leveraging Health Information Exchange to Improve Population Health Reporting Processes: Lessons in Using a Collaborative-Participatory Design Process. *EGEMS*. 2014;2(3):1082. doi:10.13063/2327-9214.1082.

10. Pilemalm S, Timpka T. Third generation participatory design in health informatics-Making user participation applicable to large-scale information system projects. *J Biomed Inform*. 2008;41(2):327-339. doi:10.1016/j.jbi.2007.09.004.

11. Van Velsen L, Illario M, Jansen-Kosterink S, et al. A Community-Based, Technology-Supported Health Service for Detecting and Preventing Frailty among Older Adults: A Participatory Design Development Process. *J Aging Res*. 2015;2015:216084. doi:10.1155/2015/216084.

12. Kangovi S, Grande D, Carter T, et al. The use of participatory action research to design a patient-centered community health worker care transitions intervention. *Healthc (Amsterdam, Netherlands)*. 2014;2(2):136-144. doi:10.1016/j.hjdsi.2014.02.001.

13. Gordon M, Henderson R, Holmes JH, Wolters MK, Bennett IM. Participatory design of ehealth solutions for women from vulnerable populations with perinatal depression. *J Am Med Informatics Assoc*. 2015;ahead of p. doi:http://dx.doi.org/10.1093/jamia/ocv109.

14. Lidstrom H, Lindskog-Wallander M, Arnemo E. Using a Participatory Action Research Design to Develop an Application Together with Young Adults with Spina Bifida. *Stud Health Technol Inform*. 2015;217:189-194.

15. Stephens KA, Lin C-P, Baldwin L-M, et al. LC Data QUEST: A Technical Architecture for Community Federated Clinical Data Sharing. *AMIA Summits Transl Sci Proc*. 2012;2012:57. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3392065&tool=pmcentrez&rendertype=abstract\nhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC3392065/.

16. Stephens KA, Lin C-P, Baldwin L-M, Echo-Hawk A, Keppel GA. A web-based tool for cataloging primary care electronic medical record federated data: FInDiT. In: *CTSA 2011 Informatics Annual Meeting*. Bethesda, MD; 2011.

17. Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res*. 2005;15(9):1277-1288. doi:10.1177/1049732305276687.

18. Trauth E, Jessup L. Understanding computer-mediated discussions: positivist and interpretive analyses of group support system use. *MIs Q*. 2000;24(1):43-79. doi:10.2307/3250979.

19. Rubio DM, Schoenbaum EE, Lee LS, et al. Defining translational research: implications for training. *Acad Med*. 2010;85(3):470-475. doi:10.1097/ACM.0b013e3181ccd618.