Journal of the American Medical Informatics Association, 00(0), 2021, 1–10
doi: 10.1093/jamia/ocab073
Research and Applications

Research and Applications

Identifying and prioritizing benefits and risks of using privacy-enhancing software through participatory design: a nominal group technique study with patients living with chronic conditions

Theodoros V. Giannouchos,[1,2] Alva O. Ferdinand,[3,4] Gurudev Ilangovan,[1] Eric Ragan,[5] W. Benjamin Nowell,[6] Hye-Chung Kum,[1,3] and Cason D. Schmit[1,3]

[1]Population Informatics Lab, School of Public Health, Texas A&M University, College Station, Texas, USA, [2]Pharmacotherapy Outcomes Research Center, College of Pharmacy, University of Utah, Salt Lake City, Utah, USA, [3]Department of Health Policy and Management, School of Public Health, Texas A&M University, College Station, Texas, USA, [4]Southwest Rural Health Research Center, School of Public Health, Texas A&M University, College Station, Texas, USA, [5]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, Florida, USA, and [6]Patient-Centered Research, Global Healthy Living Foundation, Upper Nyack, New York, USA

Corresponding Author: Cason D. Schmit, JD, Department of Health Policy and Management, School of Public Health, Texas A&M University, 212 Adriance Lab Rd, College Station, TX 77843, USA; schmit@tamu.edu

Received 20 November 2020; Revised 1 March 2021; Accepted 29 March 2021

ABSTRACT

Objective: While patients often contribute data for research, they want researchers to protect their data. As part of a participatory design of privacy-enhancing software, this study explored patients' perceptions of privacy protection in research using their healthcare data.

Materials and Methods: We conducted 4 focus groups with 27 patients on privacy-enhancing software using the nominal group technique. We provided participants with an open source software prototype to demonstrate privacy-enhancing features and elicit privacy concerns. Participants generated ideas on benefits, risks, and needed additional information. Following a thematic analysis of the results, we deployed an online questionnaire to identify consensus across all 4 groups. Participants were asked to rank-order benefits and risks. Themes around "needed additional information" were rated by perceived importance on a 5-point Likert scale.

Results: Participants considered "allowance for minimum disclosure" and "comprehensive privacy protection that is not currently available" as the most important benefits when using the privacy-enhancing prototype software. The most concerning perceived risks were "additional checks needed beyond the software to ensure privacy protection" and the "potential of misuse by authorized users." Participants indicated a desire for additional information with 6 of the 11 themes receiving a median participant rating of "very necessary" and rated "information on the data custodian" as "essential."

Conclusions: Patients recognize not only the benefits of privacy-enhancing software, but also inherent risks. Patients desire information about how their data are used and protected. Effective patient engagement, communication, and transparency in research may improve patients' comfort levels, alleviate patients' concerns, and thus promote ethical research.

Key words: Privacy, patient perception, Nominal group technique, record linkage, patient matching

INTRODUCTION

Digital health records enable more sophisticated comparative effectiveness research (CER), but these advancements come with new privacy concerns. Digital health records allow researchers to link data about people that are stored in different systems and collected for different purposes.[1–3] The process of linking person-level records is referred to as "record linkage" or "patient matching." Historically, this process involved the disclosure of all patient identifiers contained in the converging datasets to a trusted party for accurate record linkage.[2,3] While there are substantial social benefits to linking records for research, there are also understandable privacy concerns with disclosing the needed identifiers.[4–8] Failure to address these concerns can lead to legal, ethical, and public opinion consequences that impede the potential collective benefit from health information technology investment.[8–12]

Unfortunately, understanding and addressing these privacy concerns is not straightforward. Evidence shows that patients' priorities can be contradictory. Patients demonstrate a strong interest in facilitating research while demonstrating a strong preference for stringent privacy protection. Moreover, patients' preference for privacy does not always translate into privacy-promoting behavior.[13] This latter phenomenon is known as the privacy paradox.[13]

This privacy paradox has several important implications, but 2 are relevant to this study. First, the discrepancy between a person's preference for privacy and their actual behavior pushes the onus on data custodians and data users to be ethical stewards of personal information and to effectuate the protections desired by the public. Second, public engagement is necessary to ensure that software is developed and used with appropriate consideration of the public's stated privacy concerns. The former consideration supports the use of privacy-by-design principles. The latter consideration necessitates public engagement when developing privacy-enhancing software and technology. Together, these considerations form a strong argument for a participatory privacy-by-design approach for privacy-enhancing software. In this study, we provide a case study on how to engage patients when designing record linkage software for retrospective database research with privacy in mind.

Protecting privacy and confidentiality in retrospective data analysis is complex and requires a holistic approach involving technology, statistics, governance, and a shift in culture of information accountability through transparency rather than secrecy.[6,14,15] Transparency in health research using personal data requires effective communication between patients and researchers about privacy concerns and mitigating steps.[16] However, communication between data custodians, analysts, developers, and patients should flow both ways. Privacy by design is most valuable when it effectuates the protections desired by the population at risk.[17] Thus, it is critical to engage patients when designing privacy-enhancing software and technology.

Patient engagement provides a richer understanding of patients' perspectives when designing software. For CER software and systems (eg, record linkage software), patient engagement can help mitigate potential risks, build trust in CER, and build capacity for continuous research that balances legal responsibilities, patient-identified priorities, and societal benefits. Nevertheless, communicating the features of privacy-enhanced software to relevant stakeholders (eg, lawyers, institutional review boards, and patients) is an important yet difficult task. Education, communication, and engagement with the patients about these new privacy-enhancing software technologies is important because patients not only benefit from CER outcomes, but also bear the associated privacy risks.

In particular, the patient perspective is important for key stakeholders to know. Informaticians must understand and better incorporate patient privacy preferences in the balancing of benefits and risks when designing these new technologies. For data custodians to be good stewards of personal information, software systems should incorporate privacy protection principles from the beginning of the development process to deployment, use, and ultimate disposal (ie, privacy by design).[18]

While research has explored patient perspectives on the risks and benefits of using medical records for research, there is a critical gap in the literature as it relates to patient perspectives on the technical protections (ie, software) used by researchers.[19–22] Specifically, it is critical to incorporate the patient perspective when designing privacy-enhancing software so that the software can accommodate patients' desired protection into the software and system design.[21,23–27] Importantly, different patient groups might have unique perspectives for different types of software systems. For example, patients with chronic diseases are important stakeholders of record-linkage software because these patients are likely to have multiple healthcare providers with distinct electronic health records requiring linkage for scientific study.

In this study, we aim to demonstrate how patient engagement can enhance the privacy-by-design process for a novel record linkage software called MiNDFIRL (MInimum Necessary Disclosure For Interactive Record Linkage). Additionally, this study aims to identify what information about software and retrospective database research is important to patients, and by extension, for transparent research. Record linkage bears unique privacy risks because identifiers are required to accurately link records from different existing databases. This study builds on previous work examining privacy-enhancing technology for record linkage using participatory design. The ultimate outcome of this work is an open source prototype software and a set of companion documents (ie, a frequently asked questions [FAQ] template, a template data use agreement [DUA], and a template Institutional Review Board [IRB] application) to facilitate communication about the software to the different stakeholders (eg, patients, IRBs, privacy officers).[28]

Objective

This study contributes to the existing evidence on patient perspectives on privacy and research and, importantly, fills a critical gap in the literature by identifying patient priorities for important risks and benefits for privacy-enhancing software all with a goal of promoting transparency in health research using large databases. Our aim was to assess patients' perspectives on (1) the benefits of using MiNDFIRL, a privacy-enhancing software for record-linkage, (2) the potential risks that might still remain despite the use of such software, and (3) additional information on research with healthcare data that patients would like to know.[28–32] Given the importance of communication and active collaboration between patients and researchers, the findings of this research can inform guidelines for priorities of informing the public about patient data being used in research. This research will also be useful to researchers who seek to balance patients' privacy interests with anticipated societal benefits, to develop optimal and acceptable risk strategies, and to integrate patient voice in CER, improving communication practices among researchers, healthcare entities, and patients.

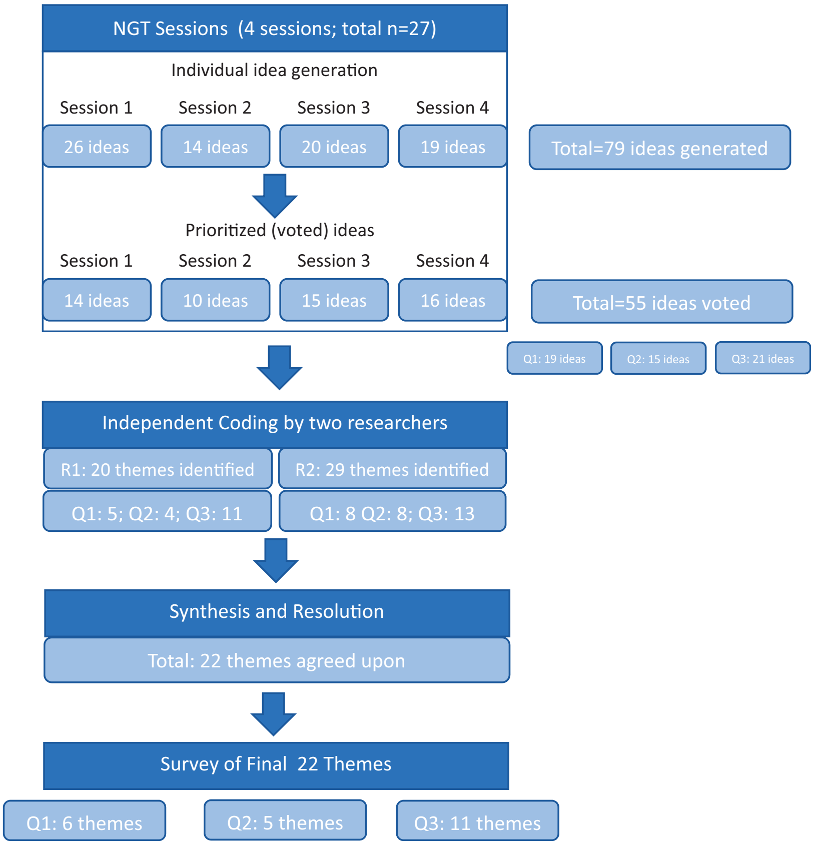


Figure 1. Idea generation and building consensus through nominal group technique sessions and online survey.

MATERIALS AND METHODS

Design

We used the nominal group technique (NGT) to solicit patient input on a generic study that used record linkage software to match patient healthcare records.[33,34] The NGT method is designed to elicit ideas from a group by mitigating the effects of dominating members in the discussions, promoting equal input from participants, and allowing assessment of the importance of the ideas generated.[34,35]

We conducted 4 NGT sessions in April 2018 with a total of 27 participants. In each NGT session, we asked participants to generate and vote on ideas in response to 3 questions related to the benefits, risks, and additional information they would like to know when their health data are used for research that requires record linkage.[33,34] Each NGT session had groups of 6 to 7 participants, within the optimal size of 5 to 12 participants recommended for an NGT to ensure the generation of broad and diverse ideas and to reach saturation.[36] After the 4 sessions, we deployed an online questionnaire (May 2018) in which themes from all 4 sessions were consolidated (Figure 1).

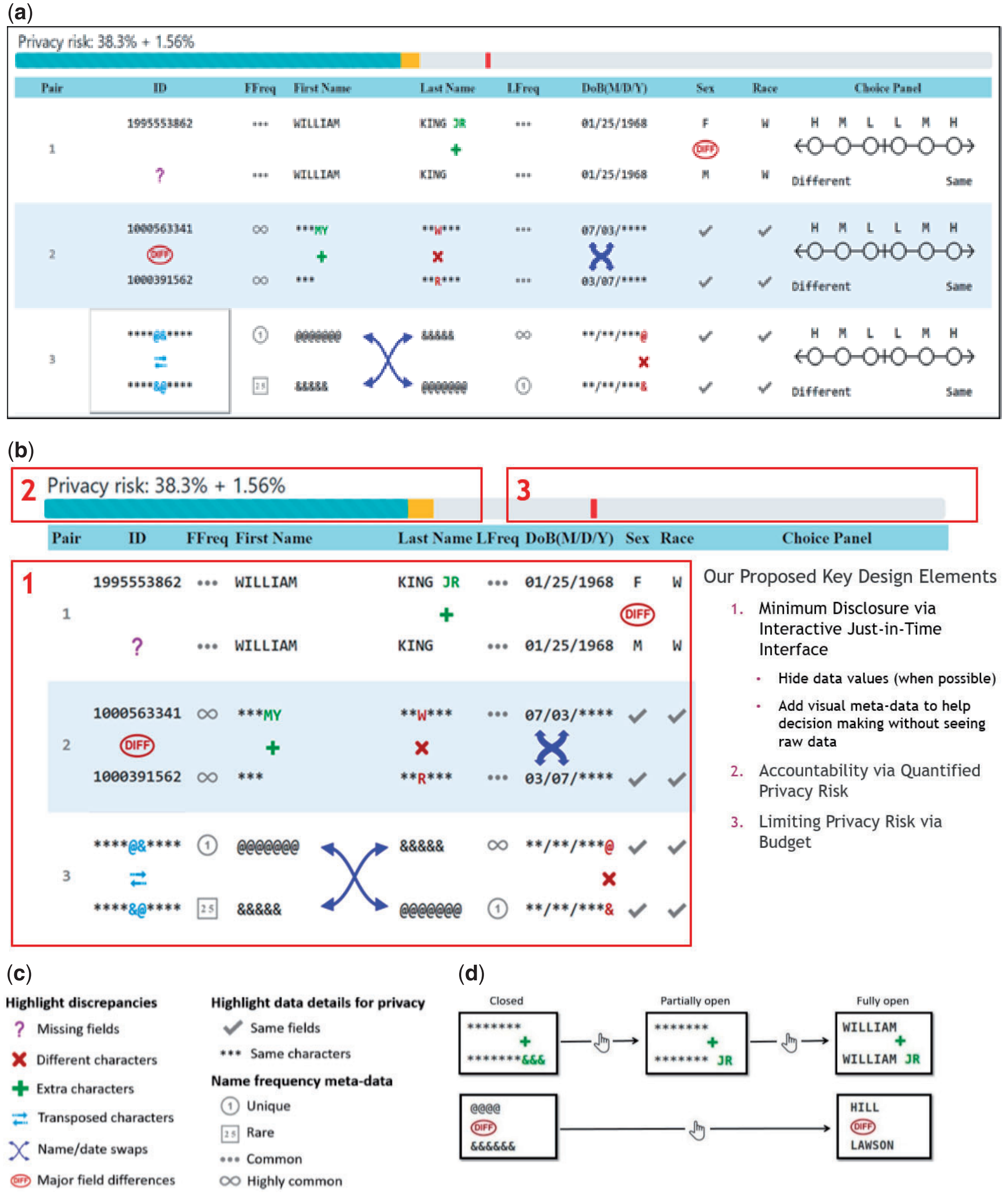
MiNDFIRL: Patient-matching prototype software

We provided participants with background information on record linkage generally, and then we gave them access to a recently designed privacy-enhancing open source software prototype (MiNDFIRL*)*, in which they experienced the record linkage process by actually using the software to link records across databases.[28–32] This software supports maximum privacy protection while still providing access to needed information to support high-quality record linkage based on (1) incremental disclosure of identifiers on an "as-needed basis," (2) minimum disclosure via recoding, (3) transparent accountability by quantifying the privacy risk due to the disclosure made, (4) limiting data access via a budget, and (5) separating the identifiers from sensitive data.[28–32] Figure 2 depicts the details of how the MiNDFIRL interactive record linkage software works. This introductory information and hands-on record linkage experience using the software was designed to enrich real-world idea generation among participants and promote research credibility.

Participants and recruitment

The study was approved by the Texas A&M University IRB. Participants were eligible to participate if they were at least 18 years of age, fluent in the English language, had at least 1 chronic condition, and had more than 3 visits to a healthcare provider within the previous year. Patients with chronic conditions are important stakeholders in retrospective database research using health data for several reasons. They have experience with personal healthcare records and related privacy issues through their regular interaction with the healthcare system, and their chronic conditions are more likely to be studied using retrospective database analyses.

We recruited patients using email listservs from 5 Patient-Powered Research Networks (PPRNs) and listervs of employees and staff at a large university. The 6 PPRNs were ArthritisPower, COPD PPRN, Health eHeart, Interactive Autism Network, Mood Network, and PRIDEnet, which are part of the national Patient-Centered Clinical Research Network. Our main goal was to include patients with diverse conditions and demographics to the extent possible. Participants received a $40 gift card at the end of the NGT session as compensation for their participation. Each participant also entered a raffle to receive an additional $100 gift card after completing the final combined online survey.



Our Proposed Key Design Elements

1. Minimum Disclosure via Interactive Just-in-Time Interface
   - Hide data values (when possible)
   - Add visual meta-data to help decision making without seeing raw data
2. Accountability via Quantified Privacy Risk
3. Limiting Privacy Risk via Budget

Highlight discrepancies
- ? Missing fields
- ✗ Different characters
- + Extra characters
- ⇄ Transposed characters
- ╳ Name/date swaps
- DIFF Major field differences

Highlight data details for privacy
- ✓ Same fields
- ••• Same characters

Name frequency meta-data
- ① Unique
- 25 Rare
- ••• Common
- ∞ Highly common

Figure 2. (A) During the tutorial, participants select an option in the "Choice Panel" on whether the data in each pair correspond to the same or different people. If they need more information to decide, participants "click" on different parts of the identifying information to see more before they decide. In essence, the visual interface for interactive record linkage masks data values and uses icons and color coding to highlight discrepancies in data pairs. Users can interactively reveal additional data details, but each access event has a "cost" that detracts from a "privacy budget."[30] (B) The main visual interface for interactive record linkage masks data values decomposed. (C) Visual masking icons used to highlight discrepancies, including matching values, and providing metadata.[32] (D) Interactive on-demand interface. Cells start with no disclosure, then partially open with a click. Cells open fully with either 1 or 2 clicks depending on the nature of the data.[30]

A total of 27 patients participated in the NGT sessions. On average, participants' mean age was 48 ± 15.5 years, they had a chronic condition for around 14 ±13.7 years, and made around 5 ± 2.4 visits to their physician during the previous year (Table 2). The 3 most frequently reported chronic conditions were chronic obstructive pulmonary disease, mental health, and high blood pressure, with 19% of the participants reporting multiple chronic conditions.

Individual NGT sessions

Three NGT sessions were conducted on an online platform to allow nationwide patient-participant representation, and 1 session was conducted in person in a computer lab. We piloted initial draft questions with several patients (not included in the study) to improve the face, content, and construct validity. We then revised the questions to incorporate pilot patients' feedback. The final questions were:

1. Are there things you like about the software that you would tell your neighbors?
2. Are there concerning things about the software that you would tell your neighbors?
3. What more would you like to know?

Each session started with a brief study description and informed consent. A member of our team with legal, privacy, and bio-ethics expertise led all NGT sessions. Two additional team members provided technical support and audio-recorded the studies. One additional researcher provided administrative support and recorded.

Figure 1 shows a summary of the NGT process and the derivation of themes from participants' ideas. In all 4 sessions, participants were given access to an online document with the 3 questions and predefined sections to enter their responses. Each NGT session followed a predetermined structure, which was disclosed to participants prior to the beginning of each session and was completed in approximately 120 minutes. In the first phase, the researchers reviewed the 20-minute online tutorial where participants conducted record linkage using the software in an attempt to clarify and address any questions. Subsequently, participants were given 10 minutes per question (30 minutes total) to individually generate ideas on each question and type them into an online document. Afterward, the facilitator reviewed all ideas and led open discussions for each question among the group. Common ideas were combined by the participants into broader themes. Participants were then asked to vote on the 2 most important themes per question.

Final themes generated across all sessions and online questionnaire

A final online questionnaire was used to identify consensus across all 4 NGT session participants. Two researchers independently conducted inductive thematic analyses on the results from each session to identify the common themes across all sessions. Themes that did not receive participant votes in any individual sessions were excluded from the final list of themes included in the final survey. The final list of themes was then created based on consensus among the 2 researchers (Figure 1). Two other researchers validated the final list by matching the participant-provided ideas with one of the final themes.

In the final online questionnaire, we asked all participants to rank-order all themes in this list. The themes from the benefits and risk questions were ranked by significance. The themes related to additional information were rated on a 5-point Likert scale to indicate the level of necessity to disclose the information to patients (eg, as an FAQ list for a research project using the software).

RESULTS

Across all 4 sessions, participants generated a total of 79 ideas in response to the 3 elicitation questions (Figure 1). Of those ideas, 24 did not receive any votes in the final phase of the sessions, resulting in 55 ideas for thematic analysis. Table 1 displays examples of participant responses and corresponding themes. For example, "researcher conscientiousness" was suggested as a benefit in 2 separate sessions, but it did not receive votes in either session and was thus dropped from thematic analysis. No ideas were dropped from the risk question. Throughout the process, similarities and common themes shared across the 4 groups provided an indication of saturation.

Of the 19 ideas related to what participants liked about MiNDFIRL, the concepts of "minimum disclosure" or limiting data accessible on an "as-needed basis" were indicated as benefits of the software in all 4 sessions. Similarly, multiple participant-generated ideas indicated that they saw to "enhanced" or "comprehensive" privacy protection through improved software support as an important benefit.

Participants also raised and voted for 15 ideas related to lingering privacy or software concerns. Most of the ideas fell under 3 themes: (1) remaining potential for some information disclosure, (2) remaining potential for information misuse by authorized users, and (3) remaining potential for hacking by unauthorized users. These themes emerged in 3 of the 4 sessions.

Participants provided 21 ideas for additional information that they would want to know about the use of the software for record linkage. Seven of the 21 generated ideas were related to the people and organizations doing the record linkage, including information on researcher identities, qualifications, personnel training requirements, and background checks. These ideas were combined into a broader theme about custodianship and control of patient data.

Using this inductive process, participant voting on the 55 generated ideas resulted in 22 final themes. Table 1 shows these final themes with contributing responses from participants. The final survey revealed the most important themes. Among the 6 choices related to the benefits of using the MiNDFIRL software, participants considered the software's "allowance for minimum disclosure" and for "comprehensive privacy protection that is not currently available" as the most important benefits with mean rank of 4.9 of 6 and 4.3 of 6, respectively (a rank of 6 indicates most important) (Figure 3A). Participants were generally most concerned about the "required checks to ensure privacy protection" and the "potential of misuse by authorized users" (eg, negligence, malfeasance) with mean scores of 3.5 of 5 and 3.4 of 5 respectively in the final vote (a rank of 5 indicates most important) (Figure 3B). Among the 11 additional information choices that participants deemed necessary to include in the FAQ, the median responses were "essential" for 1 choice, "very necessary" for 6 choices, and "necessary" for the remaining 4 choices, indicating that participants were interested in details of the study in the FAQ beyond the benefits and risks (Figure 3C).

DISCUSSION

By themselves, software developers are poorly suited to fully design usable software systems. Experts encourage the use of iterative cycles of stakeholder engagement, design, development, and testing. This engagement helps developers align software design options with varied stakeholder preferences and priorities (eg, researchers, patients).[37,38]

In the case of privacy-enhancing software, understanding the perspective of the data subjects (eg, patients) is critical because the privacy-enhancing design options are predominantly for their benefit. Moreover, because individuals' privacy preferences often conflict with their actual behaviors (ie, the privacy paradox), understanding which privacy-enhancing design options are most critical cannot be accomplished without proper engagement.[13] This engagement not

Table 1. Inductive thematic analyses of the responses and voted ideas through the 4 nominal group technique sessions.

Q1: Are there things you like about the software that you would tell your neighbors?

A. Software allows for minimum disclosure; identifiers can be opened on an as-needed basis

"Minimum disclosure—data would be accessible on an 'as needed' basis"; "Minimal disclosure of information"; "Minimum disclosure—only open as needed"; "Ability to get more detailed information as needed"; "Minimum disclosure on an as needed basis"

B. Software allows for comprehensive privacy protection that is not available now

"Better quality privacy protections"; "Comprehensive protection"; "Software adds protection to what is out there now"; "Privacy above utility"; "Identity theft could potentially be mitigated because not all information is shared"; "Blinding and privacy";

C. Software allows for participants to feel good about the use of their data in a safe manner while still having confidence in the quality of the results

"Enables safer participation"; "User security; being able to participate and still feeling good about it"; "Patients can expect that privacy is a consideration that the researchers have in mind"

D. Software allows for better accuracy in the record linkage process and the study results

"Framework allows for better accuracy in the record linkage process in that it allows users to click symbols only as needed"; "Better accuracy in the record linkage process"; "Allows for participants to have more confidence in quality of the data and the study"

E. Software is configurable to optimize safe data use per project

"The software allows for configuration on a project-to-project basis"

F. Software allows for tracking disclosures to enhance accountability

"Disclosures can be tracked and accountability can be enhanced"

Q2: Are there concerning things about the software that you would tell your neighbors?

A. Still requires checks and balances beyond the software to ensure protection (eg, accountability for software configuration, checking for secure system setup)

"Still requires checks and balances beyond the software to ensure protection (eg, accountability for software configuration, checking for secure system setup) and requirements are needed for safeguards beyond the software"

B. Still potential for misuse of information by authorized users (eg, negligence, not sufficient training)

"Misuse of information: eg, hacking, dishonest personnel. Concerns about identity theft"; "Potential misuse for future research"; "The software is not the area of concerns, but the concern may remain with human elements"; "Ulterior motives may still be at play"; "There is still the possibility of unqualified individuals doing the record linkage process"

C. Still potential for some information disclosure which may lead to false sense of protection

"There is still some a possibility for information disclosure"; "Some information disclosure is needed to ensure the accuracy of record linkages"; "Potential for false sense of privacy protection"; "Checks and balances over the parameters of what can be disclosed during the study"

D. Still potential for hacking (ie, misuse of information by unauthorized users)

"Potential for hacking"; "Misuse of information: eg, hacking, dishonest personnel; Concerns about identity theft"; "Software may not be fail proof"; "What safeguards have been put in place?"

E. Still potential for errors in the linkage process

"Learning curve for those doing the linkage"; "Still room for errors"

Q3: How necessary is it to include the following items in a frequently asked questions (FAQ) webpage for a research project using the software?

A. Who is the data custodian of the linked data (ie, who has control of the data), where is the linkage taking place (ie, which organization) and who will be doing it

"Who is doing the linkage and seeing the data? eg, background checks, training, are they following a protocol"; "Institutions or entities that will be conducting the linkage"; "The number of people that will see the data"; "The person that is analyzing the data is accountable"; "Are the record linkage personnel trained?"; "Training, experience, and re-certification information about the researchers and whether training differs based on the tasks of individual researchers on the team"; "Is there trusted third party oversight such as background checks"

B. What is the purpose and scope of the study, including how the data will be used after the linkage?

"Purpose of study"; "Explicitly list why you need these identifiers and the purpose of the research"; "How is data used (How is the data used after the linkage)"

C. What accountability mechanisms (eg, background checks, training, protocols) exist for persons involved in the research?

"The person that is analyzing the data is accountable"; "Who is doing the linkage and seeing the data? eg, background checks, training, are they following a protocol"

D. Why are identifiers needed for this research?

"Explicitly list why you need these identifiers and the purpose of the research"; "A bulk of their information will be blinded and only as many identifiers as needed to make sure that records are linked correctly will be disclosed."

E. What infrastructure is in place to safeguard the data?

"Infrastructure for keeping data secure"; "What kind of infrastructure will the data be stored in?"; "What security systems and validations are in place to prevent hacking, to receive data and to reduce 'passerby disclosure'"

F. Where can I get more information?

"Do you have a patient mentor/navigator?"

G. Will the linked data be used for other purposes?

"Statement about who owns the data and whether the raw data is non-transferable"; "Unless they give further consent, the data used to conduct the record linkage will not be used for other purposes"

H. What is the protocol in the case of misuse?

"Legal ramifications and accountability for misuse (malicious intent and negligence)"

(continued)

I. What other information, besides personal identifiers, are used during linkage?
"Is family information available to facilitate linkage decisions?"
J. How will results be disseminated?
"Statement on how dissemination of results will unfold"
K. Has the software been used before for research and has it enhanced protection as well as improve research quality?
"Actual practice in larger scale using software"

Listed are the resulting 22 themes (denoted with alphabetical labels) organized under the 3 elicitation questions (Q1-Q3) along with examples of contributing participant responses for each theme.

only leads to better software, but also participatory software design can importantly increase transparency and trust in privacy-enhancing software.

This article fills an important gap in the literature by identifying benefits and risks of privacy-enhancing software that are important to patients, as well as identifying what information about the software is most important to disclose to them. Moreover, this article provides a case study for using participatory design in software development.

The results from this study informed the design of the prototype MiNDFIRL software and its companion documents, specifically an FAQ to improve patient communication and transparency about MiNDFiRL.[16,39] The findings on the most important perceived benefits and risks as well as additional information patients want to know informed the creation of our FAQ document in patient voice. Although this study sought patient feedback on a prototype MiNDFIRL software for record linkage, we expect that other software developers will find our qualitative results and our participatory approach informative.

Participants reported that features that minimize identifier disclosure (ie, "as needed" disclosure and initial identifier masking) were the most important benefits of MiNDFIRL. Consistent with our results, previous work found that patients' opinions toward acceptability and willingness to share their information are critically influenced by the disclosure or nondisclosure of certain information, particularly social security numbers, insurance details, and names.[20,21,40] While patients almost unanimously consent to the use of anonymous data, some are reluctant to share sensitive information when identifiers are not removed.[20,40,41] In addition, transparency efforts (eg, additional information and education programs) that demonstrate the benefits of using such data have proven critical to alleviate concerns and improve uptake of promoting access and linkage of healthcare data.[42] This indicates patients' desire to support research provided that safeguards for privacy protection of their information are in place. Hence, steps to increase privacy protections and mitigate information privacy concerns are valued by the patient community.[41]

However, patients in our study had lingering concerns regarding the inherent risks of using personal identifiable information in public health and secondary databases research that require record linkage and data sharing. Our study participants were mostly concerned about additional provisions to ensure protection beyond the software and potential misuse of information by authorized users. In particular, they mentioned that fear of misuse of information and identity theft driven by ulterior motives and by dishonest or unqualified personnel is a remaining risk in CER that involves record linkage, despite the features of privacy-enhancing software. This finding suggests that user accountability and software functionality that facilitates audits (eg, proper logging) are critically important privacy-enhancing features to patients. It also suggests that patients' trust in researchers is fragile and underlines the importance of transparent participatory design for privacy-enhancing software.

This concern of authorized user malfeasance extends results from previous work that indicate patients' increased concern about data breaches and improper access by unauthorized users (ie, hackers).[21,43] Concern for unauthorized users was ranked fourth in our study, below the concern for authorized user malfeasance. Such concerns are negatively associated with consent for health information disclosure and thus highly correlated with the privacy regulatory framework that governs such studies.[43]

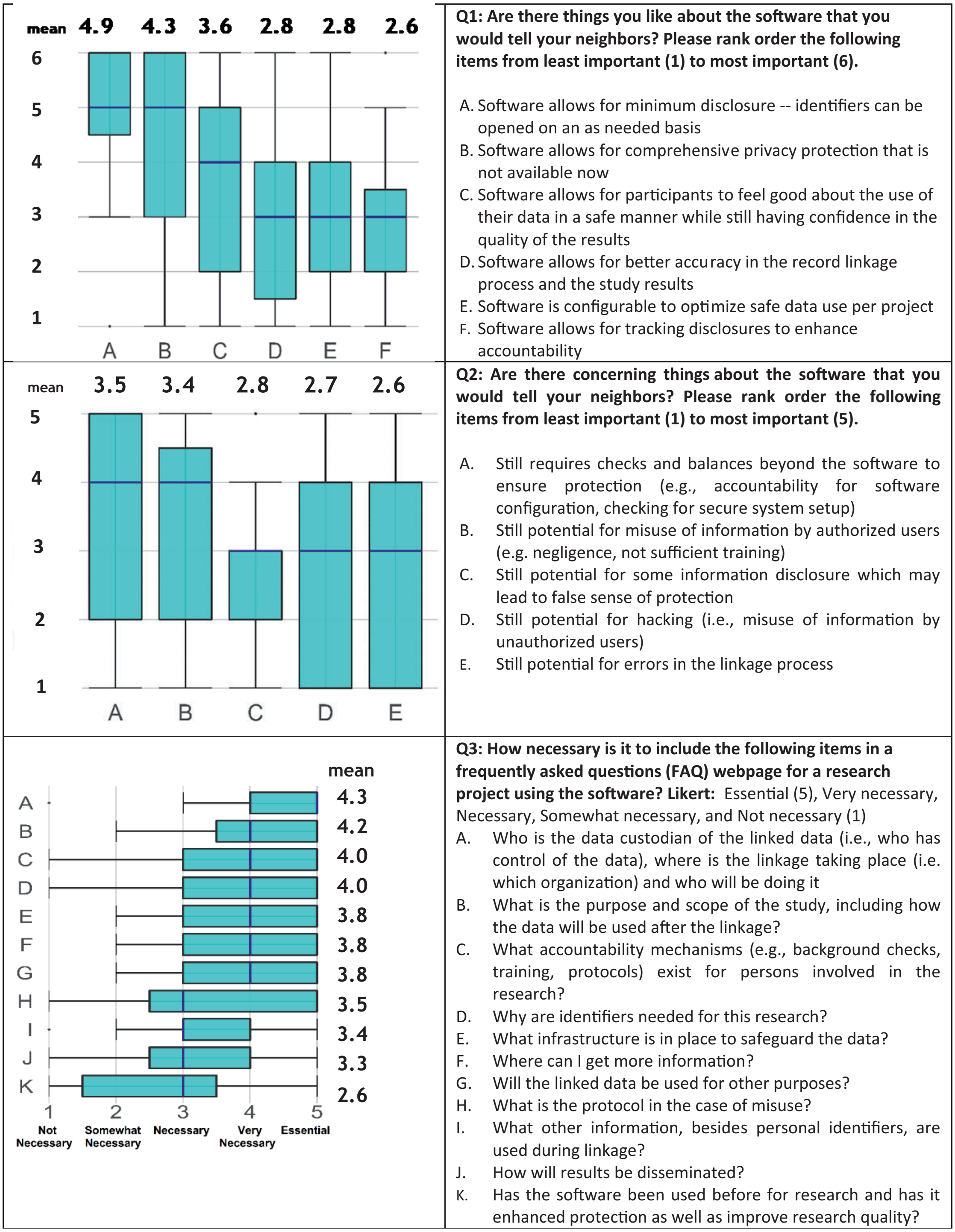
Participants identified additional information that might mitigate their concerns about future studies that might use privacy-enhancing software such as MiNDFIRL. Participants predominantly wanted information on the previous risks related to misuse, the entity that controls the data, the study's purpose, the reasons for using identifiers, and the accountability and the safeguard protocols. The requested additional information suggests that our participants were willing to rationalize their decision making through patient-oriented guidelines. Hence, our results support the notion that patients will agree to disclose their personal information if the perceived benefits of disclosure are higher than the perceived risks, as long as they are well informed about the safeguards in place to negate such risks.[44]

Our results also support the notion that increased transparency and effective communication with individuals can decrease resistance to the research use of their personal data. Individuals' intentions and follow-up actions are positively affected by the potential gain of disclosure and negatively affected by their expected loss from potential privacy violation.[45] The acceptability of sharing information is also influenced by the nature of the data recipient.[9,20] Previous work found that more than 70% of patients were comfortable sharing their data for research when informed that data access was limited to authorized personnel and security measures were in place, which was almost twice as much patients who were not explicitly informed about authorization and security measures.[20] This is consistent with our findings regarding the additional information that patients would like to know (eg, data custodian, study purpose, accountability measures). Hence, more attention should be directed toward transparency and communication efforts addressing administrative controls (eg, institutional policies, ethical oversight, research practices) to address patients' concerns. Additionally our findings suggest that privacy is also an organizational issue. Organizations without policies governing appropriate use of personal information will likely face increased friction and resistance for data sharing.[45]

Because patients bear the risks of sharing personal and sensitive information, their engagement in health technology design and development is important and necessary.[17,20,46] Our study highlights the positive impact of involving and partnering with patients in early-phase data and information technology projects (such as the design of the MiNDFIRL software) in an attempt to increase satisfaction and to obtain quality- and safety-related feedback.[20,47] Involvement and feedback from patients that reflect the diversity of the patient community can allow researchers to gain a better understanding of the diversity in patient needs.[47,48]

Limitations

Our study is not without limitations. Although we were able to recruit a sufficient number of patients consistent with the recom-



Q1: Are there things you like about the software that you would tell your neighbors? Please rank order the following items from least important (1) to most important (6).

A. Software allows for minimum disclosure -- identifiers can be opened on an as needed basis
B. Software allows for comprehensive privacy protection that is not available now
C. Software allows for participants to feel good about the use of their data in a safe manner while still having confidence in the quality of the results
D. Software allows for better accuracy in the record linkage process and the study results
E. Software is configurable to optimize safe data use per project
F. Software allows for tracking disclosures to enhance accountability

Q2: Are there concerning things about the software that you would tell your neighbors? Please rank order the following items from least important (1) to most important (5).

A. Still requires checks and balances beyond the software to ensure protection (e.g., accountability for software configuration, checking for secure system setup)
B. Still potential for misuse of information by authorized users (e.g. negligence, not sufficient training)
C. Still potential for some information disclosure which may lead to false sense of protection
D. Still potential for hacking (i.e., misuse of information by unauthorized users)
E. Still potential for errors in the linkage process

Q3: How necessary is it to include the following items in a frequently asked questions (FAQ) webpage for a research project using the software? Likert: Essential (5), Very necessary, Necessary, Somewhat necessary, and Not necessary (1)

A. Who is the data custodian of the linked data (i.e., who has control of the data), where is the linkage taking place (i.e. which organization) and who will be doing it
B. What is the purpose and scope of the study, including how the data will be used after the linkage?
C. What accountability mechanisms (e.g., background checks, training, protocols) exist for persons involved in the research?
D. Why are identifiers needed for this research?
E. What infrastructure is in place to safeguard the data?
F. Where can I get more information?
G. Will the linked data be used for other purposes?
H. What is the protocol in the case of misuse?
I. What other information, besides personal identifiers, are used during linkage?
J. How will results be disseminated?
K. Has the software been used before for research and has it enhanced protection as well as improve research quality?

Figure 3. (A) Ranked benefits from final online survey with "6" indicating most significant (top panel). (B) Ranked risks from final online survey with "5" indicating most significant (middle panel). (C) Additional information desired by participants from final online survey (bottom panel).

mended requirements for NGT sessions and we achieved saturation of themes, we note that most of our participants were primarily non-Hispanic White, and were educated, commercially insured, living with at least 1 chronic condition, and interacting with multiple healthcare providers per year. Consequently, our results might not be fully representative of the privacy concerns and generalizable to the overall U.S. population, particularly because health status directly or indirectly affects information disclosure intentions.[49] In addition, owing to the qualitative nature of our study, we were not able to conduct any stratified analyses to assess potential differences in participants' viewpoints. However, we note that one of the ultimate goals of this study was to determine the most relevant information for members of the public to understand data usage because such information could be presented in an FAQ. Thus, this study

Table 2. Sociodemographic and clinical characteristics of participants (N = 27).

Years with chronic condition(s)	14.2 ± 13.7
5 or less	30
6-10	26
11-15	15
16 or more	30
Number of physician visits	4.7 ± 2.4
Top chronic conditions	
COPD	19
Mental health	19
High blood pressure	11
High cholesterol	7
Irritable bowel syndrome	7
Lung condition	7
Thyroid	7
Leukemia	4
Long QT syndrome	4
Asthma	4
Digestive issues due to cancer treatment	4
Renal failure	4
Congestive heart failure	4
Atrial fibrillation	4
Diabetes	4
Insurance coverage	
Private	59
Medicare	15
Medicaid	4
Dual (Medicare and Medicaid)	11
VA or DoD	4
Other	7
Age, y	48 ± 15.5
Age group	
19-44 y	33
45-64 y	48
>65 y	19
Sex	
Female	63
Male	37
Race/ethnicity	
Non-Hispanic White	74
Non-Hispanic Black	7
Hispanic	7
Non-Hispanic Asian/Pacific Islander	11
Income groups	
<$25 000	15
$25 000-$75 000	59
$75 000-$125 000	15
More than $125 000	11
Education level	
High school graduate or equivalent (GED)	7
Some college	11
College graduate	56
More than college	26

Values are mean ± SD or %.

COPD: chronic obstructive pulmonary disease; DoD: Department of Defense; VA: Veterans Affairs.

was only the first step in gathering patient input.[16] Finally, although we provided educational information and practical exposure to our software before facilitating the study, some responses might have still been influenced by the limited familiarity of participants with healthcare technology and record linkage and pertinent only to MiNDFIRL software features.

CONCLUSION

The paradoxical misalignment between a person's privacy preferences and their actual behavior underlines the importance of privacy-enhancing software. It is incumbent on software developers to effectuate the privacy protections that patients value. Moreover, comprehensive privacy-enhancing software is the key to handling the complicated balance between benefits and risks of using healthcare records for public health and research.

Our study is one of the first in the literature to engage patients to actively learn about and experience how well-designed software may protect privacy while still allowing benefits of using individual data for public health. There was clear consensus among the patients in the benefits of using privacy-enhancing software for comprehensive protection while still having confidence in research quality. Yet, risks that are inherent in record linkage processes still remain—particularly those related to accountability, misuse of data, and data accuracy. Nonetheless, additional information provided to patients by researchers might alleviate patients' concerns and promote future research with important implications at the community level. Our study highlights the importance of active and effective patient engagement to advance linkage methods, transparency, and acceptable risk strategies to improve patients' comfort levels with data sharing.

FUNDING

Research reported in this article was funded in part through a Patient-Centered Outcomes Research Institute Award (ME-1602-34486). The statements presented in this article are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute, its Board of Governors, or its Methodology Committee.

AUTHOR CONTRIBUTIONS

H-CK conceived this study. H-CK, AOF, CDS, TVG, and WBN designed the study. H-CK, AOF, CDS, TVG, and GI facilitated the sessions, collected the data, and conducted the data analysis in consultation with EDR. CDS, TVG, AOF, and H-CK led the writing of this manuscript with all authors critically reviewing and providing feedback on subsequent drafts. H-CK, EDR, and GI designed the study software. All authors gave their approval for the final version to be submitted and published.

CONFLICT OF INTEREST STATEMENT

The authors declare not conflict of interest.

DATA AVAILABILITY STATEMENT

Some of data underlying this article cannot be shared publicly to protect the privacy of individuals that participated in the study. Other data will be shared on reasonable request to the corresponding author.

REFERENCES

1. Setoguchi S, Zhu Y, Jalbert JJ, *et al.* Validity of deterministic record linkage using multiple indirect personal identifiers linking a large registry to claims data. *Circ Cardiovasc Qual Outcomes* 2014; 7: 475–80.

2. Kelman CW, Bass AJ, Holman CDJ. Research use of linked health data - a best practice protocol. *Aust N Z J Public Health* 2002; 26: 251–5.
3. Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. *Annu Rev Public Health* 2011; 32: 91–108.
4. Ricciardi TN, Lieberman MI, Kahn MG, *et al.* Clinical terminology support for a national ambulatory practice outcomes research network. *AMIA Annu Symp Proc* 2005; 2005: 629–33.
5. Thompson TG, Brailer DJ. *The Decade of Health Information Technology: Delivering Consumer-Centric and Information-Rich Health Care Framework for Strategic Action*. Washington, DC: U.S. Department of Health and Human Services, Office of the National Coordinator for Health Information Technology; 2004.
6. Kum H-CC, Krishnamurthy A, Machanavajjhala A, *et al.* Social genome: putting big data to work for population informatics. *Computer* 2014; 47: 56–63.
7. Hansson MG, Lochmüller H, Riess O, *et al.* The risk of re-identification versus the need to identify individuals in rare disease research. *Eur J Hum Genet* 2016; 24: 1553–8.
8. Kaye J. The tension between data sharing and the protection of privacy in genomics research. *Annu Rev Genom Hum Genet* 2012; 13: 415–31.
9. Whiddett R, Hunter I, Engelbrecht J, *et al.* Patients' attitudes towards sharing their health information. *Int J Med Inform* 2006; 75: 530–41.
10. Xu H, Dinev T, Smith J, *et al.* Information privacy concerns: Linking individual perceptions with institutional privacy assurances. *J Assoc Inform Syst* 2011; 12: 798–824.
11. Malhotra NK, Kim SS, Agarwal J. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Inf Syst Res* 2004; 15: 336–55.
12. Hulkower R, Penn M, Schmit C. Privacy and confidentiality of public health information. In: Magnuson JA, Dixon BE, eds. *Public Health Informatics and Information Systems*. 3rd ed. London, United Kingdom: Springer; 2020: 147–66.
13. Barth S, de Jong MDT. The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review. *Telemat Inform* 2017; 34: 1038–58.
14. Wartenberg D, Thompson WD. Privacy versus public health: The impact of current confidentiality rules. *Am J Public Health* 2010; 100: 407–12.
15. Narayanan A, Shmatikov V. Myths and fallacies of personally identifiable information. *Commun ACM* 2010; 53: 24–6.
16. Schmit C, Ajayi K. V, Ferdinand AO, *et al.* Communicating with patients about software for enhancing privacy in secondary database research involving record linkage: Delphi study. *J Med Internet Res* 2020; 22: e20783.
17. Årsand E, Demiris G. User-centered methods for designing patient-centric self-help tools. *Inform Health Social Care* 2008; 33: 158–69.
18. Shapiro SS. Privacy by design: moving from art to practice. *Commun ACM* 2010; 53: 27–9.
19. Damschroder LJ, Pritts JL, Neblo MA, *et al.* Patients, privacy and trust: Patients' willingness to allow researchers to access their medical records. *Soc Sci Med* 2007; 64: 223–35.
20. Kass NE, Natowicz MR, Hull SC, *et al.* The use of medical records in research: what do patients want? *J Law Med Ethics* 2003; 31: 429–33.
21. O'Brien EC, Rodriguez AM, Kum HC, *et al.* Patient perspectives on the linkage of health data for research: Insights from an online patient community questionnaire. *Int J Med Inform* 2019; 127: 9–17.
22. Kaufman DJ, Murphy-Bollinger J, Scott J, *et al.* Public opinion about the importance of privacy in biobank research. *Am J Hum Genet* 2009; 85: 643–54.
23. Fleurence R, Selby J. V, Odom-Walker K, *et al.* How the patient-centered outcomes research institute is engaging patients and others in shaping its research agenda. *Health Aff (Millwood)* 2013; 32: 393–400.
24. Chudyk AM, Waldman C, Horrill T, *et al.* Models and frameworks of patient engagement in health services research: A scoping review protocol. *Res Involv Engagem* 2018; 4: 28.
25. Patient-Centered Outcomes Research Institute. https://www.pcori.org/. Accessed March 2, 2020.
26. Turvey CL, Klein DM, Nazi KM, *et al.* Racial differences in patient consent policy preferences for electronic health information exchange. *J Am Med Inform Assoc* 2020; 27: 717–25.
27. Giannouchos T, Kum H-C, Ferdinand AO, *et al.* Patients' and Stakeholders' perceptions of risk and benefits of the Privacy Preserving Interactive Record Linkage (PPIRL) framework. In: *AcademyHealth Advancing Ethical Research Annual Meeting*; 2018.
28. Population Informatics Lab. MiNDFIRL. GitHub. 2020. https://github.com/pinformatics/mindfirl. Accessed February 25, 2021.
29. Kum H-C, Krishnamurthy A, Machanavajjhala A, *et al.* Privacy preserving interactive record linkage (PPIRL). *J Am Med Inform Assoc* 2014; 21: 212–20.
30. Kum H-C, Ragan ED, Ilangovan G, *et al.* Enhancing privacy through an interactive on-demand incremental information disclosure interface: applying privacy-by-design to record linkage. In: *SOUPS'19: Proceedings of Fifteenth USENIX Conference* on Usable Privacy and Security; 2019: 175–89.
31. Li Q, D'Souza AG, Schmit C, *et al.* Increasing transparent and accountable use of data by quantifying the actual privacy risk in interactive record linkage. arXiv, doi: http://arxiv.org/abs/1906.03345. 10 Jun, 2019, preprint: not peer reviewed..
32. Ragan ED, Ilangovan G, Kum HC, *et al.* Balancing privacy and information disclosure in interactive record linkage with visual masking. In: *CHI'18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*; 2018: 1–12.
33. Harvey N, Holmes CA. Nominal group technique: an effective method for obtaining group consensus. *Int J Nurs Pract* 2012; 18: 188–94.
34. Horton JN. Nominal group technique. *Anaesthesia* 1980; 35: 811–4.
35. Potter M, Hamer PW, Gordon S, *et al.* The Nominal Group Technique: a useful consensus methodology in physiotherapy research. *N Z J Physiother* 2004; 32: 126–30.
36. Bouchard TJ, Hare M. Size, performance, and potential in brainstorming groups. *J Appl Psychol* 1970; 54: 51–5.
37. Shneiderman B, Plaisant C, Cohen M, *et al. Designing the User Interface: Strategies for Effective Human-Computer Interaction*. 5th ed. Boston, MA: Pearson; 2010.
38. Paulovich B. Language design to improve the health education experience: using participatory design methods in hospitals with clinicians and patients. *Visible* 2015; 49: 145–59.
39. Population Informatics Lab. mindfirl/docs. GitHub. 2020. https://github.com/pinformatics/mindfirl/tree/master/docs. Accessed May 26, 2020.
40. Phelps J, Nowak G, Ferrell E. Privacy concerns and consumer willingness to provide personal information. *J Public Policy Mark* 2000; 19: 27–41.
41. Hann IH, Hui KL, Lee SYT, *et al.* Overcoming online information privacy concerns: an information-processing theory approach. *J Manag Inf Syst* 2007; 24: 13–42.
42. Angst CM, Agarwal R. Adoption of electronic health records in the presence of privacy concerns: The elaboration likelihood model and individual persuasion. *MIS Q* 2009; 33: 339–70.
43. Bellman S, Johnson EJ, Kobrin SJ, *et al.* International differences in information privacy concerns: a global survey of consumers. *Inf Soc* 2004; 20: 313–24. doi:10.1080/01972240490507956
44. Culnan MJ, Bies RJ. Consumer privacy: balancing economic and justice considerations. *J Soc Issues* 2003; 59: 323–42.
45. Culnan MJ, Armstrong PK. Information privacy concerns, procedural fairness, and impersonal trust: an empirical investigation. *Organ Sci* 1999; 10: 104–15. doi:10.1287/orsc.10.1.104
46. Chung AE, Vu MB, Myers K, *et al.* Crohn's and colitis foundation of America partners patient-powered research network. *Med Care* 2018; 56: S33–40. doi:10.1097/MLR.0000000000000771
47. Leung K, Lu-McLean D, Kuziemsky C, *et al.* Using patient and family engagement strategies to improve outcomes of health information technology initiatives: Scoping review. *J Med Internet Res* 2019; 21: e14683.
48. Frampton SB, Guastello S, Hoy L, *et al. Harnessing Evidence and Experience to Change Culture: A Guiding Framework for Patient and Family Engaged Care. NAM Perspectives*. Discussion paper. Washington, DC: National Academy of Medicine; 2017.
49. Zhang X, Liu S, Chen X, *et al.* Health information privacy concerns, antecedents, and information disclosure intention in online health communities. *Inf Manag* 2018; 55: 482–93.