



PATIENT-CENTERED OUTCOMES RESEARCH INSTITUTE
FINAL RESEARCH REPORT

Developing and Testing Software for Linking Patient Data from Multiple Sources

Hye-Chung Kum, PhD, MSW¹; Eric Ragan, PhD²; Alva Ferdinand, DrPH, JD¹; Cason Schmit, JD¹

AFFILIATIONS:

¹Texas A&M University, College Station

²University of Florida, Gainesville

Original Project Title: Privacy Preserving Interactive Record Linkage (PPIRL) via Information Suppression
Institution Receiving Award: Department of Health Policy and Management, Texas A&M University Health Science Center

PCORI ID: ME-1602-34486

To cite this document, please use: Kum H-C, Ragan E, Ferdinand A, Schmit C. (2022). *Developing and Testing Software for Linking Patient Data from Multiple Sources* Patient-Centered Outcomes Research Institute (PCORI).
<https://doi.org/10.25302/04.2022.ME.160234486>

TABLE OF CONTENTS

ABSTRACT	5
BACKGROUND.....	7
Record Linkage in Patient-Centered Outcomes Research/Comparative Effectiveness Research Studies	7
Figure 1. Record Linkage	7
RL Challenge	8
The Need for a Better Solution.....	10
Objective	11
Outcome: Open-Source Software and 3 Companion Documents.....	12
Figure 2. PPIRL Research Agenda	13
PARTICIPATION OF PATIENTS AND OTHER STAKEHOLDERS.....	14
Table 1. Stakeholder Information and Roles	15
Aims 1 and 2: Design and Evaluation of the MiNDFIRL Interface.....	17
Aims 3: Three Companion Documents Developed Through Participatory Action Research.....	18
METHODS AND RESULTS	20
Privacy by Design: Approximate RL Using a Hybrid Human-Computer System	20
Figure 3. Full Approximate RL Process Using a Hybrid Human-Computer System	20
Aims 1 and 2: Software Design Overview.....	21
Figure 4. Participatory Design	21
Figure 5. The Visual Interface for Interactive Record Linkage Masks Data Values and Uses Icons and Color Coding to Highlight Discrepancies in Data Pairs.....	22
Figure 6. The Main Visual Interface for Interactive Record Linkage Masks Data Values Decomposed.....	22
Figure 7. Visual Masking Icons Used to Highlight Discrepancies, Including Matching Values, and Providing Metadata	23
Figure 8. Interactive On-Demand Interface	24
Study A.1 and A.2: Formative Evaluations	26
Figure 9. Examples Showing 1 Record Pair Under the 5 Different Experimental Conditions	29
Figure 10. Differences in Percentage of Character Values Revealed With the Different Conditions Applied to the Generated Test Data Used for the Experiment.....	32

Table 2. Summary of Participant Characteristics by Condition for Study A.1	33
Figure 11. Record Linkage Accuracy for the 5 Conditions	33
Table 3. Accuracy of Linkage Decisions (% of Correct Responses) by Condition.....	34
Table 4. Completion Time by Condition	34
Table 5. Participant Confidence in Linkage Decisions by Condition (Scale, 1-3)	34
Table 6. Accuracy Effect Size	35
Table 7. Time Effect Size.....	35
Table 8. Confidence Effect Size	36
Figure 12. Visual Summary Representing the Differences of the 5 Experimental Conditions In the Evaluation	39
Table 9. Summary Characteristics by Condition for Study A.2	40
Figure 13. KAPR Privacy Scores for the 5 Conditions	41
Figure 14. Percentage of Incorrectly Linked Pairs From the 5 Conditions	41
Figure 15. Time Taken to Complete the Linkage Task for the 5 Conditions.....	42
Table 10. Error Rate of Linkage Decisions (% of Wrong Responses) by Condition.....	43
Table 11. Record Linkage Time by Condition	43
Table 12. Privacy Risk Score (KAPR) by Condition	43
Table 13. Error Effect Size	44
Table 14. Time Effect Size.....	45
Table 15. KAPR Effect Size	45
Study B: Summative Evaluation.....	46
Figure 16 (a). Data Flow for UTH and UAB Study	50
Aim 3: Develop 3 Companion Documents Through Participatory Action Research.....	54
Figure 17. Idea Generation and Building Consensus Through 4 NGT Sessions and an Online Survey	57
Table 16. Sociodemographic and Clinical Characteristics of Participants	59
Figure 18. Benefits, Risks, and Additional Information From the Final Online Survey (N = 27)	61
Figure 19. Overview of Delphi Process and Round Content.....	63
Table 17. Demographic Information of Participants Who Completed Round 1 (N = 38)	66
Table 18. Themes of Participant Feedback	68
Table 19. Sociodemographic, Clinical Characteristics, and Privacy Attitude Scores of Participants	70

Figure 20. Privacy Statement Format and Usefulness Preference	71
Figure 21. Example Quotes for Positive and Negative Feedback	74
Figure 22. Example FAQ Screenshot	75
Table 20. Emerging Themes From ELSI Experts.....	78
Figure 23. Consensus Votes of Emerging Themes From ELSI Experts	79
DISCUSSION	84
Minimum Necessary Standard and Practical Challenges	84
Synergies With Related Research in RL	85
Our Contributions to Privacy-Enhancing Technology Development for CER	86
Figure 24. Reduction in Information Disclosure While Maintaining Human Decision-Making in RL	87
Study Limitations.....	90
Future Research	92
CONCLUSIONS	94
REFERENCES	96
RELATED PUBLICATIONS.....	104
Published (Submitted in Appendix)	104
Draft Completed to Be Submitted in the Near Future	104
ACKNOWLEDGMENTS	105
APPENDIX: OUTCOME	106
MiNDFIRL (MInimum Necessary Disclosure For Interactive Record Linkage)	106

ABSTRACT

Background: Comparative effectiveness research (CER) and patient-centered outcomes research (PCOR) routinely use secondary data (eg, insurance claims, health records). Leveraging secondary data requires effective and accurate *record linkage* (RL), that is, matching the same individuals in different data sets. The absence of common, error-free, unique identifiers across data sources challenges RL and forces the use of identifying information (ie, names) to ensure proper linkage. This, in turn, raises privacy concerns. While automated methods are useful, high-quality RL requires human interaction (eg, parameter settings, building training data sets, validating results). Consequently, managing errors from imperfect and complex real-world data requires human access to identifiable data.

Objectives: Broadly, our objective was to investigate privacy-enhancing RL tools that can facilitate accurate matching with a hybrid human-computer system that strictly controls information disclosure. Specifically, we aimed to (1) design effective information visualizations for RL, (2) determine optimal levels of information disclosure for RL, and (3) develop consensus with patients and stakeholders on what they need to know about how RL is conducted. Using these findings, the main outcomes were to design (1) prototype open-source software, (2) a template privacy statement, (3) an IRB application template, and (4) a template data use agreement to share information about the software with appropriate stakeholders.

Methods: This research used methods from 2 fields. First, we used a human-computer interaction agile software development approach to develop the prototype software called Minimum Necessary Disclosure For Interactive Record Linkage (MiNDFIRL), including controlled user studies (N > 100), expert surveys, and case studies. Second, we used nominal group technique (NGT) focus groups and Delphi studies commonly used in participatory action research to engage stakeholders in the research. These methods were used to understand perceived benefits, risks, and practical concerns with the new privacy-enhancing approach that MiNDFIRL employs. Patients and ELSI (ethical, legal, and social implications), including IRB, experts were engaged to develop the 3 companion documents for MiNDFIRL. We then conducted an online survey (N > 400) to obtain public opinion of the developed privacy statement.

Results: For iterative software design and development, the project includes multiple formative evaluations through (i) 2 controlled experiments with volunteer nonexpert participants and (ii) an expert review. The first experiment (study A.1: N = 104) evaluated human decision-making in RL with the visual data-masking technique. A second experiment (study A.2.1: N = 122) focused on the on-demand interactive interface design for incrementally disclosing partial information. Collectively, the results demonstrate the ability to greatly limit the amount of identifying information available to human decision makers (only 7.85% compared with 100% with all data disclosed) without negatively affecting decision quality or completion time. We also conducted an expert review with 6 experts (study A.2.2). Their feedback supports the notion that a level of access to identifying information that is intermediate between “all or nothing” can provide better accuracy than that with no access but more protection than with full access. As a

summative evaluation, 2 case studies were conducted to evaluate our approach in more realistic and complex operational scenarios at (i) the University of Texas Health Science Center at Houston (UTH; study B.1) and (ii) the University of Alabama at Birmingham Health System (UAB; study B.2). The studies consisted of RL with electronic health records (N = 10 000 total pairs with 303 manually reviewed pairs) and patient-generated data (N = 1055 total pairs, with 187 manually reviewed pairs), respectively. Both the UTH and UAB results demonstrate that the default disclosure budget for identifying information in MiNDFIRL, at 30%, based on results from the formative studies, was sufficient for most human decision-making in RL. Our engagement research to develop template companion documents for the MiNDFIRL software included 4 studies: an NGT session with 11 ELSI experts (study D.1), an NGT session with 27 patients (study C.1), a Delphi study with 13 ELSI experts (study D.2), and a Delphi study with 33 patients (study C.2). Generally, we identified consensus across all studies. The potential to reduce risk to the minimum necessary was a main perceived benefit of our approach, while concerns still remained for needed organizational administrative controls (eg, software configuration, and secure system setup) across all studies. In a nationally representative sample (study C.3: N = 470), more than 80% were satisfied with the privacy statement that was developed in a web-based, interactive, frequently asked questions format.

Conclusions: Our controlled experiments demonstrate that properly designed software can enhance privacy while supporting legitimate access for human decision-making. The results also suggest limits to how much data can be hidden before negatively influencing the quality of decisions. We also found that public privacy statements, written to reflect patients' voices and interests, can increase transparency and improve patient trust. Based on these findings, we designed, implemented, and released the open-source MiNDFIRL prototype software along with 3 companion documents describing the use of the software for high-quality RL to support CER/PCOR.

Limitations: The current prototype software code, MiNDFIRL, needs to be fully developed for use across CER/PCOR. Additionally, the project scope did not include investigating automated algorithms required for a comprehensive hybrid human-computer system. Although we observed thematic saturation from the respondents, our qualitative studies (ie, NGT and Delphi) might not broadly reflect the full range of divergent opinions of all groups. Nonetheless, our large-scale, nationally representative sample did not find any differential preferences across socioeconomic status, providing support for the cocreated frequently asked questions language.

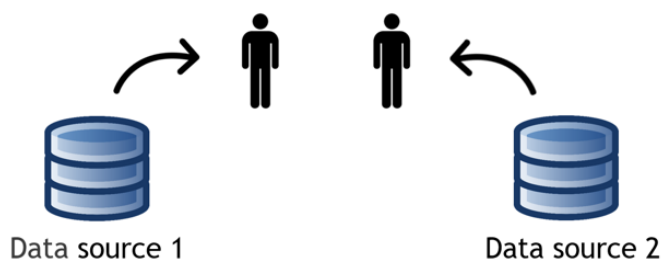
BACKGROUND

Record Linkage in Patient-Centered Outcomes Research/Comparative Effectiveness Research Studies

Population data at the individual level can inform patient-centered outcomes research (PCOR)/comparative effectiveness research (CER) studies. However, such data are often collected through different systems, resulting in separate, unconnected databases. Linking data from disparate databases has clear benefits, such as enabling treatment improvements, comparative effectiveness studies of different interventions (eg, policies, treatments), and decision-making for investing limited resources to improve health. For example, 2 National Institute on Drug Abuse studies integrated data from multiple state agency databases on 56 923 Medicaid beneficiaries with opioid dependency to conclude that buprenorphine was cheaper and safer than alternative treatments.^{1,2} Analyzing linked cancer registries and Medicare and Medicaid data has yielded significant results describing the patterns and outcomes of cancer care, including health disparities.³⁻⁶

Integrating data from diverse, heterogeneous systems is essential to properly leverage available data, but this requires an effective means of linking records. The goal of *record linkage* (RL) is to match rows that represent the same real-world entity (eg, different records pertaining to the same person) in multiple databases or in the same database (eg, deduplication) (Figure 1).⁷ Accurate RL is critical for building data research networks and replicable science.

Figure 1. Record Linkage



Record linkage is to match rows that represent the same real-world entity (eg, different records pertaining to the same person) in multiple databases or in the same database (eg, deduplication).

While automated RL methods can help integrate many records from heterogeneous data systems, high-quality RL requires human interaction to manage the inevitable errors and discrepancies resulting from imperfect and complex real-world data.^{8,9} For example, humans can often determine whether slight differences between 2 records suggest that the records belong to twins or to the same person who has a nickname in 1 record. Errors that are not properly managed in machine-only data integration systems propagate to subsequent data analyses, which can lead to potential problems with invalid results and poor decision-making. Thus, researchers need a means of providing direct control over the RL process to limit and bound errors. However, because of the personal and often sensitive nature of human data, privacy becomes a serious concern. The goal of this research is to investigate secure techniques and develop effective tools to accurately integrate data from heterogeneous sources while still protecting confidentiality by using a hybrid human-computer system that securely and strictly controls identifying information disclosure during RL.

RL Challenge

Three key RL challenges are (1) the lack of common, error-free, unique identifiers (eg, medical record numbers, names) across data sources; (2) the necessity of human involvement for high-quality RL; and (3) the need to use patient-level identifying information to ensure proper linkage. These challenges limit both the quantity and quality of studies using linked data, as well as their reproducibility. Addressing these challenges is essential to meet the growing need for studies that use linked data from large data repositories.

RL Challenge: No Common, Error-Free Unique Identifier

The absence of a common, error-free, unique identifier makes exact matching solutions insufficient because many correct matches will be missed due to minor and immaterial differences (eg, misspellings, use of nicknames, change in last name due to marriage). Approximate methods (probabilistic or deterministic) are necessary alternatives, but they require data cleaning, standardization, and manual resolution of ambiguous matches.^{7,10-13} Probabilistic approximate methods generate a statistical probability score that 2 records from different sources represent the same entity, based on a model developed from the data. In

comparison, deterministic approximate matching methods are rule based, where the researcher specifies the rules under which the 2 records are considered a match (eg, all identifiers match), ambiguous (eg, some identifiers match), or a nonmatch (eg, no identifiers match).

Although efficient, machine-only RL is problematic because it may lead to selection bias as a result of preferentially matching patients with complete information on required identifiers.¹² This can underrepresent particular groups, including socioeconomically disadvantaged and racial/ethnic minorities.^{12,14-16} For example, Bronstein et al found that when matching Medicaid data to vital records, the resulting matched analytic data sets tended to underrepresent the outcomes of high-risk pregnancies.¹² Baldi and colleagues found that the covariates in Cox regression models can be biased due to not capturing all true links when analyzing survival rate in a cohort of patients with breast cancer.¹⁴ Systematic linkage errors are inevitable in automatic algorithms and can result in biases.^{12,14-16}

RL Challenge: Necessity of Human Interaction

Without a common, error-free, unique identifier, human involvement is essential to obtain high-quality, bias-free linkages. Human interaction is necessary to tune these results from machine-only systems.^{9,17,18} This necessarily means that some identifying information must be revealed to trusted persons to produce accurate linkages. In the intensive manual process, linkage experts spend months using the software to clean and tune the linkage models, during which many choices and assumptions are made. For example, when cancer registry data were linked to Medicare and Medicaid data in Michigan, of a total of 109 925 individuals who were being linked, 16 288 (15%) were confirmed through manual verification.⁴ These steps are typically difficult to document and verify. As a result, most linkages are not reproducible, because the tuning step, conducted by human experts based on human judgment, is challenging to replicate.

RL Challenge: Protecting Confidentiality

The disclosure of certain identifiers necessary for high-quality RL raises understandable confidentiality concerns. These concerns can be addressed through administrative, physical, and technical controls. When these confidentiality concerns are not appropriately addressed, there can be serious consequences for RL projects. For example, data use agreements (DUAs) might be drafted to restrict RL or severely limit the identifiers that are available for RL. Similarly, database owners may refuse to share data if their confidentiality concerns are not addressed appropriately. The absence of a trusting relationship between parties can be a potent data-sharing barrier that takes substantial time and effort to address. Parties can employ mechanisms that promote transparency and accountability as a way to address confidentiality concerns, but that sometimes can impede otherwise-permitted data uses.

The Need for a Better Solution

We posit that the direction of PCOR/CER points to many more studies that will require linking data from diverse sources. This will necessarily require that a larger pool of individuals, with varying levels of expertise, participate in RL. We believe this will require a software system that can be used by nonexperts to resolve the ambiguous matches resulting from automatic RL, clean and standardize messy data, and provide a documented tuning of linkage rules to ensure reproducibility, all in a way that protects confidentiality.

The gap in the existing methods and software is demonstrated well in the challenges faced by national Health Information Exchange (HIE) efforts.¹⁹ Jim Younkin, director of Pennsylvania's HIE, noted that 25% of master patient index (MPI) records did not contain valid information in key identifying fields (eg, name and birthdate). The quality of MPI varies widely, and most MPIs have duplicate records that must be cleaned during RL.²⁰ Dev Culver, director of Maine's HIE, believes that duplicate record rates are as high as 3% because of an aversion bias against incorrect links. Consequently, they err on the side of not linking for ambiguous linkages, so true links are missed, leaving databases fragmented.¹⁹

For 2 decades, government institutions have called for new RL methods. A 2001 US Government Accountability Office report on RL identified a need for linking person-level data while citing the importance of properly handling identifying data.²¹ More recently, an Institute of Medicine (IOM; now the National Academy of Medicine) report titled, “Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research”²² [HIPAA, Health Insurance Portability and Accountability Act] gave recommendations on conducting data-intensive health research. The IOM recommendations included creating an RL mechanism “so that more useful data sets can be made available for research in a manner that protects privacy, confidentiality, and security.”²² In 2013, the Office of the National Coordinator for Health Information Technology identified accurate RL “as a significant challenge for the past decade,” arguing that “patient safety is the driving force” for improving RL, and urged support for an open-source algorithm to test RL accuracy.²³ In the same year, the PCORI Methodology Committee report also reflects the importance of transparent, reproducible RL tools.²⁴ A 2014 Agency for Healthcare Research and Quality report concluded with calling for more research on secure and accurate RL tools, given the importance of RL activities for enhancing observational CER.¹⁵ In sum, all major leaders in Health Information Technology^{15,21-24} have stated the importance and challenges of privacy and RL.

This research investigates a novel interactive software interface that presents users with fully masked, deidentified data, but it allows users to reveal more information if required to make good decisions.

Objective

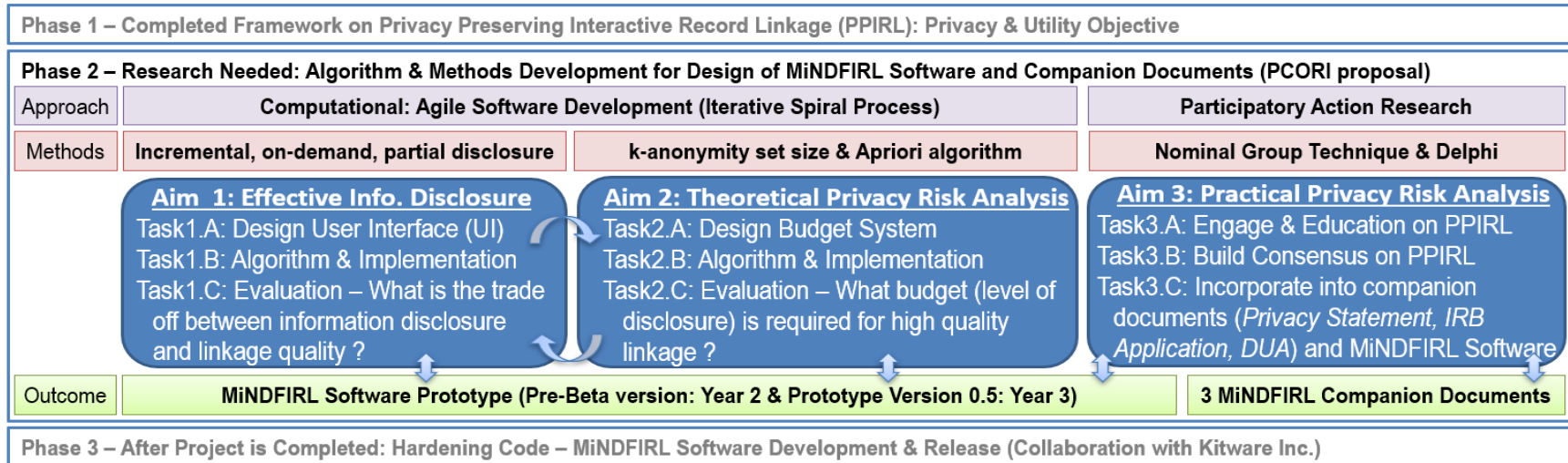
Broadly, our objective was to investigate privacy-enhancing RL tools that can facilitate accurate matching using a hybrid human-computer system that strictly controls information disclosure. Specifically, we aimed to (1) design effective information visualizations for RL, (2) determine optimal levels of information disclosure for RL, and (3) develop consensus with patients and stakeholders on what they need to know about how to build public trust in secondary database research studies where informed individual consent is not possible (Figure 2). Thus, the scope of this project and report is to expand knowledge of and enhance

privacy in manual (ie, human-driven) RL processes through mixed-methods and software user studies. The scope of this research is not intended to expand the existing knowledge of automated RL processes. We have used automated RL methods only to create appropriate record samples to evaluate human-computer interactions for manual RL. More details on our automated RL research can be found in previously published work.^{25,26} Accordingly, this research adheres to the general standards of the field of human-computer interaction. Figure 2 summarizes the research agenda for this project.

Outcome: Open-Source Software and 3 Companion Documents

With these aims (Figure 2) in mind, the main outcomes were to design (1) a prototype open-source software called MInimum Necessary Disclosure For Interactive Record Linkage (MiNDFIRL) and 3 companion documents: (2.1) a template privacy statement, (2.2) an IRB application template, and (2.3) a template DUA to share information about the software with appropriate stakeholders. All of these products, along with an online tutorial and YouTube video, have been released as open-source code and documents on GitHub and our public project website.^{27,28}

Figure 2. PPIRL Research Agenda



Abbreviations: DUA, data use agreement; MiNDFIRL, MInimum Necessary Disclosure For Interactive Record Linkage; PPIRL, privacy-preserving interactive record linkage.

PARTICIPATION OF PATIENTS AND OTHER STAKEHOLDERS

We engaged patients and stakeholders throughout the project. We obtained feedback and guidance through active participation in our studies, regular meetings with our advisory committees (ACs; methods committee and a user committee), as well as quantitative and qualitative research (on-site and online). Patient and stakeholder engagement was utilized to increase the real-world impact and relevance of this work.²⁹

As a methods project, patient and stakeholder engagement may not follow the more-conventional CER projects typically funded by PCORI. Instead, we identified and engaged appropriate stakeholder groups for the development of our RL methodology (aims 1 and 2) and the companion documents (aim 3). In sum, the key stakeholders identified as necessary to increase real-world impact on developing the RL software and companion documents were (1) individuals with technical methods expertise in RL, user interface design, and open-source software development (Table 1, “RL and SW methods expertise” column); (2) users of RL software for CER (Table 1, “User of RL SW for CER” column); (3) ethical, legal, and social implications (ELSI; including IRB) experts familiar with data governance and security issues in RL (Table 1, “ELSI expert” column); and (4) patients whose data are being used in RL. Stakeholders were engaged in all 3 aims as partners in study design as well as participation in periodic AC meetings and as study participants.

Table 1 provides the full list of key stakeholders and their roles. One important part of engagement was the 20-minute online tutorial³⁰ on MiNDFIRL and general RL that we developed for use in all of our individual studies across all aims. The tutorial was key to getting meaningful feedback about RL and the MiNDFIRL interface from the participants. In total, over 800 study participants, detailed below under each aim, from the general public living in the United States completed the online tutorial³⁰ on MiNDFIRL and general RL. We note that the feedback gathered through all patient and stakeholder engagement efforts was the main source of information for our final software design and was incorporated as appropriate. The input and contribution of patients and stakeholders were crucial for us to design, develop, evaluate, finalize, and validate our software and the accompanying documents. Ultimately,

Table 1. Stakeholder Information and Roles

Name	Affiliation	Expertise	RL and SW methods expertise	User of RL SW for CER	ELSI expert	Patient perspective	Methods AC	User AC	Role
Jeffery Curtis	UAB	Clinical, research data network PI (user), CER, PCOR, ELSI		X	X		X	X	Study B.2 UAB research design; feedback on initial SW design and companion documents
Elmer Bernstam	UT Houston	Health informatics, MPI, CER, research data network co-PI (user)	X	X	X		X	X	Study B.1 UTH research design; feedback on initial SW design and companion documents
Ben Nowell	Great Healthy Living Foundation	ArthritisPower PPRN				X		X	Study C.1 and C.2 research design and recruitment; outreach by hosting a webinar
Sean O'Brian	Duke University	PI of PCORI project on privacy	X	X			X		Feedback on initial SW design and companion documents
Ashok Krishnamurthy	UNC at Chapel Hill	Co-I on Mid-South CDRN	X	X			X		Feedback on initial SW design and companion documents
Peter Yu	Texas A&M Univ	HIPAA privacy officer		X	X		X		Feedback on initial SW design and companion documents
Patrick Reynolds	Kitware	Open-source health SW development; user interface design	X				X		Feedback on initial SW design and companion documents

Name	Affiliation	Expertise	RL and SW methods expertise	User of RL SW for CER	ELSI expert	Patient perspective	Methods AC	User AC	Role
Daniel Basile	Texas A&M Univ	Patient (chronic illness), security officer, IT support for research			X	X	X		Feedback on initial SW design and companion documents
Michael Morrisey	Texas A&M Univ	Linking claims data		X	X			X	Feedback on initial SW design and companion documents
Eva Shipp	Texas A&M Univ	Research data network PI		X	X			X	Feedback on initial SW design and companion documents
Robin Clark	Univ of Mass	MA all-payer database		X	X			X	Feedback on initial SW design and companion documents
Alison Fraser	Univ of Utah	Linking data for cancer outcomes	X	X	X			X	Feedback on initial SW design and companion documents
Stacie Dusetzina	Vanderbilt Univ Medical Center	Pharmacoepidemiologist, HSR		X			X	X	Feedback on initial SW design and companion documents
Leonard J Nelson	Vice chair of the IRB for St. Vincent's and St. Vincent's East Hospitals	ELSI expert			X		X		Feedback on initial SW design and companion documents

Abbreviations: AC, advisory committee; CDRN, Clinical Data Research Network; CER, comparative effectiveness research; Co-I, co-investigator; ELSI, ethical, legal, and social implications; HIPAA, Health Insurance Portability and Accountability Act; HSR, health services research; IT, information technology; MPI, master patient index; PCOR: patient-centered outcomes research; PI, principal investigator; PPRN, patient-powered research network; RL, record linkage; SW, software; UAB, University of Alabama at Birmingham Health System; UNC, University of North Carolina; Univ., University; UT, University of Texas Health Science Center.

both communities had a substantial impact on the design of our software and research. Most importantly, the codeveloped companion documents can facilitate effective communication with relevant stakeholders on complex but important issues in using data for CER while addressing the concerns of both patients and stakeholders.

Aims 1 and 2: Design and Evaluation of the MiNDFIRL Interface

Specifically, the RL methods research in aims 1 and 2 supported the design and evaluation of our novel approach to achieve both high-quality RL and effective privacy protection. If successful, the prototype software may be used by researchers and professionals as a research tool to facilitate studies that involve linking data across multiple databases. Prototype software is commonly developed in computational research to evaluate designs and pilot new systems. Thus, we recruited expert advisors from a professional network of individuals conducting RL studies. We carefully selected volunteer AC members to include clinicians, methods experts, and academics based on their experience with conducting RL and in user interface design, software development, health services research, economics, privacy, and medicine. Throughout the project, we obtained feedback and recommendations on the software interface and features from these experts and stakeholders. In particular, the summative evaluation RL case studies in study B were conducted in partnership with our stakeholders from the University of Texas Health Science Center at Houston (UTH; Dr Bernstam) and the University of Alabama at Birmingham Health System (UAB; Dr Curtis) who were engaged during study design, recruitment, and interpretation.

We also conducted a formal expert survey to solicit feedback on the software design from these stakeholders. The survey solicited feedback on the MiNDFIRL software from 6 experts who regularly conduct RL and work with sensitive or identifying information (5-10 years of experience). Experts were volunteers recruited from a professional network of people conducting RL studies, including our AC members. Their feedback supports the notion that an intermediate level of access between “all or nothing” can provide better accuracy than no access but more protection than full access. This and other feedback informed the use, development, and adjustment of the software interface to make it more user-friendly for RL.

Aims 3: Three Companion Documents Developed Through Participatory Action Research

The companion documents developed in aim 3 are intended to facilitate effective communication with all stakeholders and patients on the complex but important CER issues related to RL and how MiNDFIRL is used to enhance privacy. In particular, the privacy statement is intended to communicate relevant information to the general public about the software and how projects are handling their data. Thus, we engaged the patient community extensively in aim 3 with quantitative and qualitative methods to solicit feedback and assist with codeveloping the privacy statement companion document for MiNDFIRL. Input from the patient community was essential to identify key issues that could be addressed in software design as well as to determine how to effectively communicate important issues in patient voice (ie, addressing critical interests in familiar language). Studies C.1 to C.3 (see Methods and Results section for details) engaged over 500 patients to cocreate the privacy statement template.³¹

To improve the quality and reach of our patient engagement, we worked collaboratively with Dr Ben Nowell from the ArthritisPower patient-powered research network (PPRN). He was a key partner in this process and participated in proposal writing, study design, recruitment, and outreach to all PPRNs, as well as interpretation of the results. In addition, the ArthritisPower PPRN hosted a webinar for the general patient population where we presented on key issues in RL, privacy, and the MiNDFIRL interface.

Similarly, the IRB application template companion document was intended to facilitate effective communication related to issues in human subjects research with the IRB in an RL study using MiNDFIRL. The first phase of engagement to develop this template language was to partner with ELSI experts on our advisory board for study design and feedback on initial drafts of the document and survey tools. The second phase of engagement activities involved collecting qualitative and quantitative feedback from ELSI experts and stakeholders to codevelop the IRB application template. Experts were recruited from professional networks, such as conferences and email lists, and by consulting university, hospital, and Veterans Affairs websites. We collected publicly available contact information for such individuals and reached

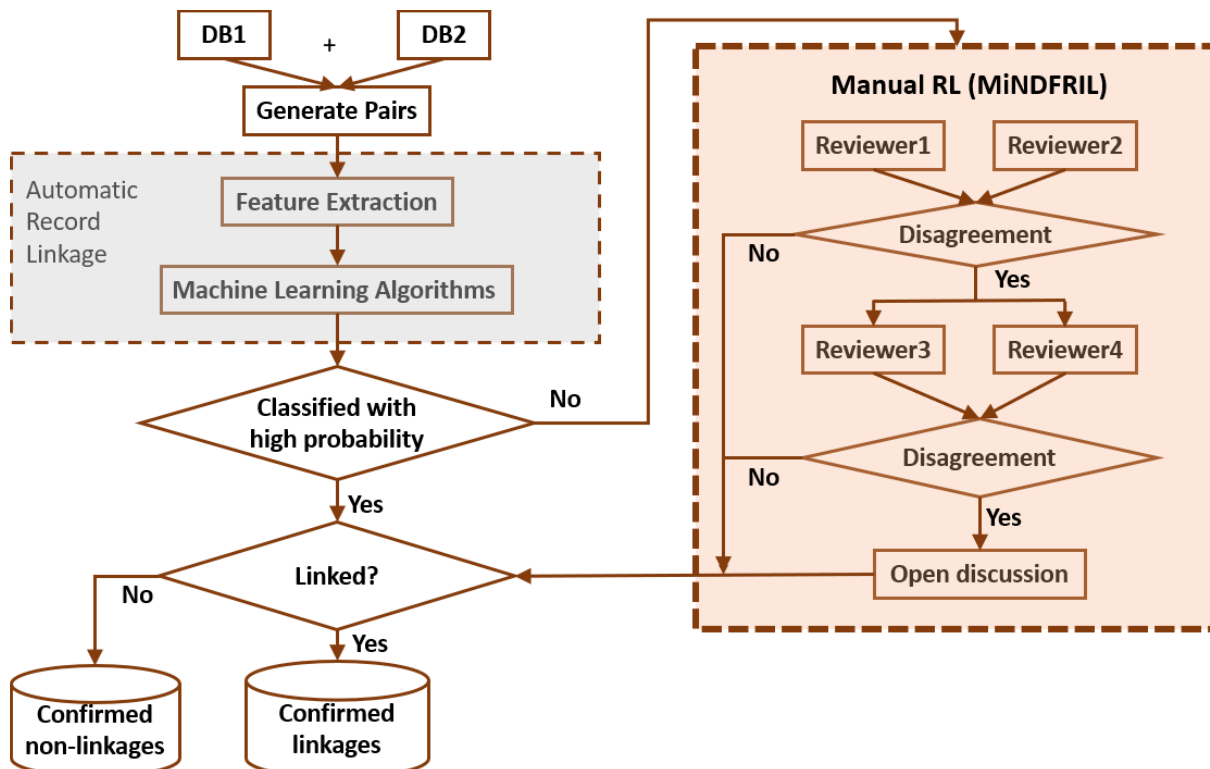
out to them via email. In our email, we informed them about our study and its purpose and timeline, and we invited them to participate. The details are discussed in the Methods and Results section under study D.1 and D.2. This feedback was essential to understanding the perceived and potential benefits and risks of MiNDFIRL, and how to communicate effectively with ethics review bodies (eg, an IRB) about the software. Engaging content experts and stakeholders ensured that our research was applicable to and useful for future researchers.

METHODS AND RESULTS

Privacy by Design: Approximate RL Using a Hybrid Human-Computer System

High-quality integration of data from several sources requires RL, and while algorithmic approaches for RL can partially automate the process, the complexity of real-world data collection and significant uncertainty in automation have necessitated augmenting these automated methods with manual human review to ensure data quality. Effective RL of data involving identifiable information often requires different people to have access to the information, which will increase personal privacy risk (eg, identity theft, data leaks, or social engineering attacks) of those whose data are stored. Figure 3 depicts the full hybrid human-computer system for approximate RL. Our research focuses on the manual RL process (the brown box) and studies the trade-offs between privacy and utility of identifying information for

Figure 3. Full Approximate RL Process Using a Hybrid Human-Computer System



Abbreviations: DB, database (data set); MiNDFIRL, Minimum Necessary Disclosure For Interactive Record Linkage; RL, record linkage.

human decision-making in RL. The 2 goals we try to balance using good software design following the privacy-by-design principle³² are as follows³³:

- **Privacy goal:** Limiting disclosure of identifying information (eg, names) and guaranteeing no disclosure of sensitive information (eg, diagnosis)
- **Utility goal:** Sustaining human effectiveness for valid RL decisions

Aims 1 and 2: Software Design Overview

The project adopted iterative processes of design, development, and evaluation utilizing specialized studies (Figure 4) to test specific elements of the developed software,³⁴ such as privacy-preserving techniques and the visual user interface. The full interface as seen by the user is shown in Figure 5. Figure 6 decomposes the interface to depict the 3 key design elements discussed below that allow MiNDFIRL to effectively implement the “minimum necessary” ethical principle for privacy protection.

Figure 4. Participatory Design

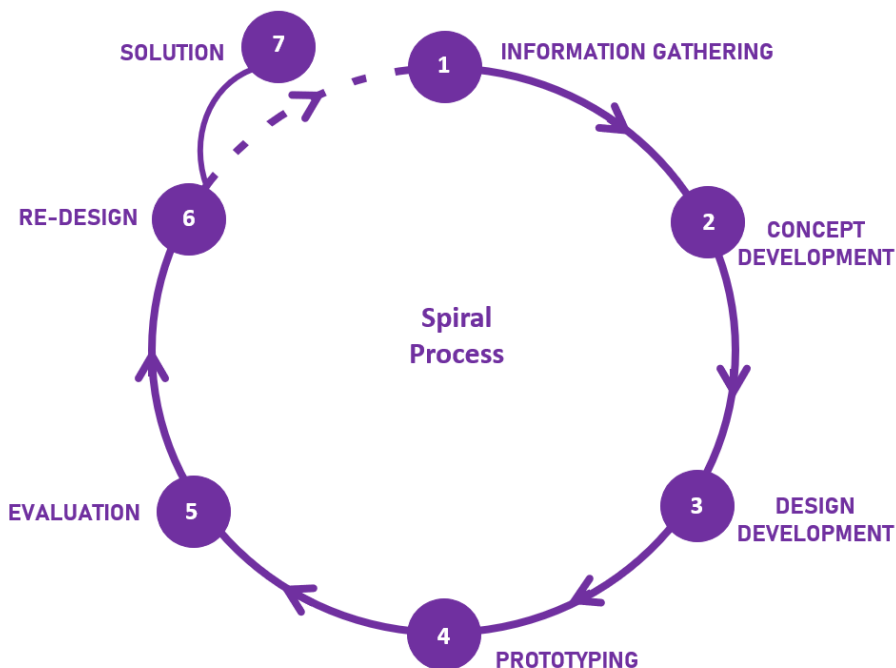
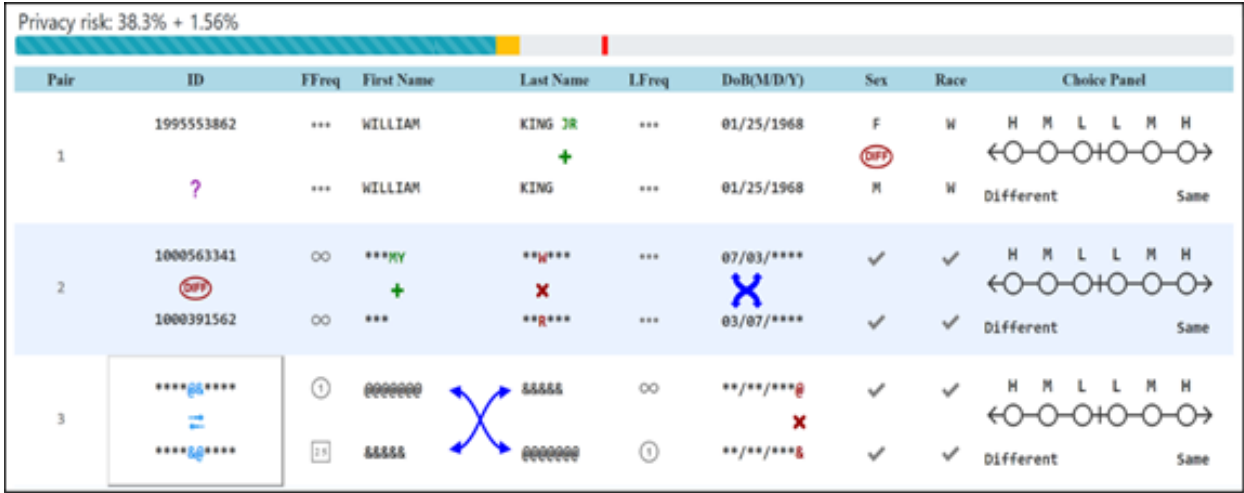


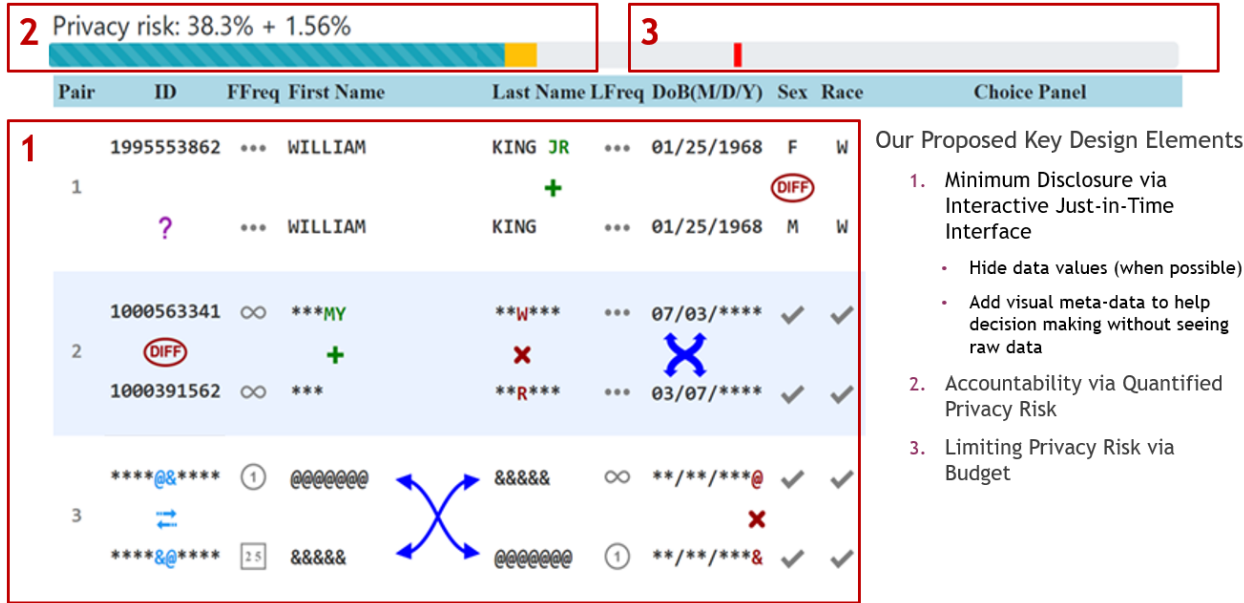
Figure 5. The Visual Interface for Interactive Record Linkage Masks Data Values and Uses Icons and Color Coding to Highlight Discrepancies in Data Pairs



Abbreviations: DoB, date of birth; FFreq, first name frequency; ID, identification; LFreq, last name frequency. Users can interactively reveal additional data details, but each access event has a “cost” that detracts from a “privacy budget.” The record linkage task requires users to decide whether the data in each pair correspond to the same or different entities.

Reproduced from Kum et al. Fifteenth USENIX Conference on Usable Privacy and Security. <https://dl.acm.org/doi/10.5555/3361476.3361489>. Reprinted with permission from SOUPS’19 (Copyright ©2019). All Rights Reserved.¹⁸

Figure 6. The Main Visual Interface for Interactive Record Linkage Masks Data Values Decomposed















Abbreviations: DoB, date of birth; FFreq, first name frequency; ID, identification; LFreq, last name frequency.

Minimum Disclosure via Interactive Just-in-Time Interface

In summary, the developed techniques and software designs manage privacy and data availability through software designed to limit the disclosure of personal details only on an “as-needed” basis while supporting accountability by recording data-access events. The method is complementary to automated linkage algorithms prior to human involvement, and the software’s graphical user interface (Figure 5) employs visual data masking to limit the amount of raw data available by default for human review. Informative icons and visual highlighting (Figure 7) are used to help users understand data discrepancies while hiding the details of the underlying identifying information. To manage the trade-offs between decision quality and data privacy, users may decide to access specific and limited data details by clicking (Figure 8) to aid decision-making for specific data discrepancies, but the software can enforce a disclosure limit (or “privacy budget”) to the total amount of raw data values capable of being accessed or revealed.

Figure 7. Visual Masking Icons Used to Highlight Discrepancies, Including Matching Values, and Providing Metadata

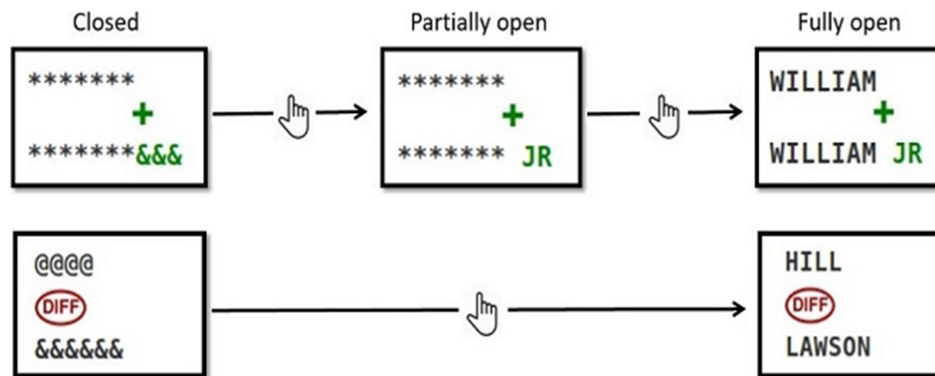
Highlight discrepancies	Highlight data details for privacy
 Missing fields	 Same fields
 Different characters	 Same characters
 Extra characters	Name frequency meta-data
 Transposed characters	 Unique
 Name/date swaps	 Rare
 Major field differences	 Common
	 Highly common

Reproduced from Kum et al. Fifteenth USENIX Conference on Usable Privacy and Security.

<https://dl.acm.org/doi/10.5555/3361476.3361489>. Reprinted with permission from SOUPS’19 (Copyright ©2019).

All Rights Reserved.¹⁸

Figure 8. Interactive On-Demand Interface



Cells start with no disclosure and then partially open with a click. Cells open fully with either 1 or 2 clicks, depending on the nature of the data.

Reproduced from Kum et al. Fifteenth USENIX Conference on Usable Privacy and Security.

<https://dl.acm.org/doi/10.5555/3361476.3361489>. Reprinted with permission from SOUPS'19 (Copyright ©2019). All Rights Reserved.¹⁸

Accountability via Quantified Privacy Risk

The proper quantification of risk is important to support transparency and the reasoning, communication, and decisions on the privacy and utility trade-off. For our system, the goal is to quantify the identity disclosure risk because sensitive attribute disclosure is fundamentally blocked by keeping the sensitive attributes separate from the identifying attributes. Thus, our prototype developed and used the k-Anonymity Privacy Risk (KAPR) score, which uses the anonymity-set size as an estimate of the identity disclosure risk.³⁵ *Anonymity-set size*, defined as the number of people in the population who share the same identifying information, is an intuitive and accessible measure to estimate the privacy risk. The larger the set size, the lower the privacy risk. For example, when a frequently occurring name (eg, Eric) is disclosed, there is a low probability that a specific person with that name could be identified. In comparison, a rare name (eg, Mahin) may be sufficient information to determine a person's identity. In addition, anonymity-set size is easily calculated dynamically for any information to be disclosed during human interaction with the system. As more information is disclosed to aid linkage, the anonymity-set size will be reduced. This in turn will increase the privacy risk.

The exact definition is given below with more details, and an example is provided in Li et al.³⁵ In sum, the KAPR score is a normalized score from 0% (nothing disclosed) to 100%

(everything disclosed), with higher scores if more is disclosed and what is disclosed is more unique (ie, increases identifiability risk). Uniqueness is defined as the number of records in the data being linked. Although the KAPR score function was used in our meter in the user study because of its accuracy for measuring identity disclosure, it is important to note that the exact function used is not as important as the use of a reasonable privacy risk feedback method that users can understand.

Definition (k-Anonymized Privacy Risk [KAPR] score). Let N be the full number of records across the databases being linked with D attributes that was used to build potential pairs for review. Let \mathfrak{X} be the information disclosure state associated with a partial display X with $2n$ records (ie, n pairs). Let p_{ij} be the proportion of characters of attribute j of record i disclosed. Let κ be the minimum allowed anonymity-set size, and let k_i be the anonymity-set size of record i based on the current disclosure state. The KAPR score is given by

$$K(\kappa; \mathfrak{X}) := \frac{\kappa}{ND} \|\mathfrak{X}\|_{1,1} = \frac{\kappa}{ND} \sum_{i=0}^{2n-1} \frac{1}{k_i} \sum_{j=0}^{D-1} |p_{ij}|.$$

Here, $\|\cdot\|_{1,1}$ is the standard $L_{1,1}$ matrix norm.

Limiting Privacy Risk via Budget

Although the interactive interface enables only the minimum necessary disclosure and the feedback meter encourages limited-access behavior to personally identifiable information (PII) and audits after the fact, neither of these designs alone can enforce limited disclosure, which may be a condition of using the data set. For example, certain DUAs may limit access to Social Security numbers (SSNs) by allowing up to 4 digits. In our system, such hard rules on data access can be enforced using an option to configure the software. MiNDFIRL, an RL software, is designed to allow configuration by a manager to control an appropriate budget for different human data workers and specific data projects. In particular, the privacy budget feature can be used to enforce a limit on the total disclosure for a given use case.

By specifying a disclosure limit or an allowable privacy budget ahead of time, the system can guarantee a certain maximum level of information disclosure for a specific RL task. Moreover, specifying a budget based on expert users can provide guidance to novice users about the right balancing point between access to data for good decisions vs trying to make matches with insufficient information which can result in lower-quality decisions.

Ultimately, the goal of any legitimate access to sensitive data is to maximize utility under a fixed privacy budget. Thus, it is important to design the system that allows for specifying the privacy budget ahead of time so that it can be enforced. Figuring out appropriate levels of privacy risk for a given task to support quality data is an open research area that will require further study. In our evaluation, we start by studying how different privacy limits might lead to different human behavior in making decisions to disclose information, as well as how these limits on the privacy score impact the quality of the RL task.

Study A.1 and A.2: Formative Evaluations

The material presented in this section previously appeared in the following 2 peer-reviewed publications:

- Study A.1 was published in Ragan E, Kum H-C, Ilangovan G, Wang H. Balancing privacy and information disclosure in interactive record linkage with visual masking. Paper presented at: 2018 CHI Conference on Human Factors in Computing Systems (CHI '18); April 21-26, 2018; Montreal, Québec, Canada. Accessed January 13, 2022. <https://dl.acm.org/doi/10.1145/3173574.3173900>.¹⁷ It won a CHI 2018 Honourable Mention Best Paper Award (top 5% of all submissions). It was also presented at the 14th Symposium on Usable Privacy and Security (SOUPS) Aug 2018 as an invited poster.
- Study A.2 was published in Kum H-C, Ragan ED, Ilangovan G, Ramezani M, Li Q, Schmit C. Enhancing privacy through an interactive on-demand incremental information disclosure interface: applying privacy-by-design to record linkage. Paper presented at: SOUPS'19: Fifteenth USENIX Conference on Usable Privacy and Security; August 11-13, 2019; Santa Clara, CA. Accessed January 13, 2022. <https://dl.acm.org/doi/10.5555/3361476.3361489>.¹⁸

Objective

As part of the participatory design process, formative studies A.1 and A.2 investigated different aspects of the MiNDFIRL software's design and user performance to inform subsequent software development decisions. The first formative study, study A.1, focused on the trade-offs in the use of visual data-masking techniques to preserve privacy vs the quality of RL decision-making when varying amounts of data are hidden from data workers. This study prioritized the evaluation of visual representations (Figure 7) to reduce the amount of record detail that is disclosed while still providing metadata to indicate the type of data errors or discrepancies between records for human RL decisions. The second formative study, study A.2, focused on evaluating the interactive just-in-time interface and the application of the developed KAPR score to enforce restrictions on information disclosure following the metaphor of a budget for allowable information access. Study A.2 investigated different budget limits and visual feedback to participants about how their data access decisions consumed their allowable budget. Both formative studies evaluated how the different design configurations influenced RL quality (accuracy of data pairs linked), efficiency (time taken to make decisions), and the amount of data details accessed (as relating to disclosure and privacy risk score). Furthermore, the formative evaluations sought to collect data about end-users' general matching strategies, behavior, and understanding of the developed methods when working with the interactive RL designs.

General Methods

As formative evaluations during the software design and testing cycle, the research project includes multiple evaluations of the core software features through 2 controlled experiments with volunteers as proxies for novice data workers. Each study allowed us to collect participant feedback and software usage data with specific design elements. The formative studies were conducted with a curated data set using derived data from a real-world data set based on targeted known data problems and challenges.¹⁷ Study sessions were conducted with a fixed procedure involving (i) an introduction to the software and linkage scenario, (ii) a period of tutorial and software use where participants conducted RL with given

sample data, and (iii) additional questions and feedback solicitation for the software features, challenges, and participants' understanding. More details can be found in the published papers attached in the Appendix.

We recruited formative study participants by sending out virtual signup sheets on multiple mailing lists at a large university for students interested in completing data science tasks. Interested participants (1) signed up for any of the sessions that they were available for, (2) indicated their student level (eg, PhD, MS, undergraduate senior, junior), and (3) indicated their home department. We collected student-level and home department information to approximately balance aptitude for the data linkage task during the random assignment process, as follows:

- Randomly assign an equal number of participants to each of the 5 experimental conditions, as specified below in the “Study design” subsection (Figure 9), using a random number generator
- Reassign to have approximately same number of graduate students in each condition
- Reassign to have approximately same number of quantitative departments defined as computer science, mathematics, management information systems, engineering, economics, and accounting. Nonquantitative departments included other liberal arts departments, basic science (eg, biology, chemistry) departments, architecture, animal science, and food science.

Based on these assignments, we sent confirmation emails for the sessions. We could only do approximate balancing because not all participants who signed up actually participated, and we also had walk-in session participants. Thus, during the actual sessions, we had to assign some participants as best possible. We collected some basic characteristics of participants during the sessions to check on the balance of factors—including age and gender—that were not included in our prior balancing efforts.

Figure 9. Examples Showing 1 Record Pair Under the 5 Different Experimental Conditions

Baseline	Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	8000002767			JUDE	WILLIAM		09/09/1906	M	W
	8000003567			JUDE	WILLIAM JR		09/09/1960	M	B
Full	Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	8000002767		⓪	JUDE	WILLIAM	⓪	09/09/1906	M	W
	8000003567		⓪	JUDE	WILLIAM JR	⓪	09/09/1960	M	B
Moderate	Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	*****27**		⓪	✓	WILLIAM	⓪	09/09/1906	M	W
	*****35**		⓪	✓	WILLIAM JR	⓪	09/09/1960	M	B
Low	Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	*****27**		⓪	✓	*****	⓪	**/**/06	M	@
	*****35**		⓪	✓	***** JR	⓪	**/**/60	M	&
Masked	Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	*****@@**		⓪	✓	*****	⓪	**/**/00	✓	@
	*****@@**		⓪	✓	*****@@@	⓪	**/**/00	✓	&

Abbreviations: DoB, date of birth; FFreq, first name frequency; ID, identification; LFreq, last name frequency. These views show the same underlying data, but the visuals and amount of symbol substitution vary based on the viewing condition.

Reproduced from Ragan et al. 2018 CHI Conference on Human Factors in Computing Systems. <https://dl.acm.org/doi/10.1145/3173574.3173900>. Reprinted with permission from CHI'18 (Copyright ©2018). Association for Computing Machinery. All Rights Reserved.

Study A.1: Static Design – Effective Visual Masking

Using an RL scenario, we conducted a controlled experiment to evaluate how various degrees of information disclosure can affect decision-making for RL tasks.

Hypotheses. This research is motivated by the need to understand the extent to which it is feasible to deidentify personal data without negatively affecting the utility of the data for decision-making. Our high-level hypothesis is that even legally deidentified data—in which personal details are hidden—can be effectively used for decision-making tasks that generally

rely on personal details, but we expect that achieving this will require an appropriate interface that can sufficiently convey the most important meta-information for the decision-making task.

Applied to the RL scenario of our study, we expect that the use of value masking and visual markup (see Figure 7) can sufficiently portray differences to limit the amount of identifying information needed to make linkage decisions. This means that we predict that records with hidden details can be linked with a level of accuracy similar to the base case with unlimited information disclosure. However, we expect to see a reduction in quality for extreme data masking because users may not have sufficient information for accurate RL when too much information is masked. We summarize these hypotheses (H1-H3) as follows:

- **H1.** With an appropriate interface, significant limits on data availability can be enforced without compromising decision quality.
- **H2.** There is a limit to how much data can be hidden before negatively influencing the quality of judgment in decisions involving person-level data.
- **H3.** The addition of supplemental visual information through masking (see Figure 7) can help expedite RL decisions by making it easier to identify the types of differences.

Study design. The experiment followed a between-subjects design with 5 conditions.

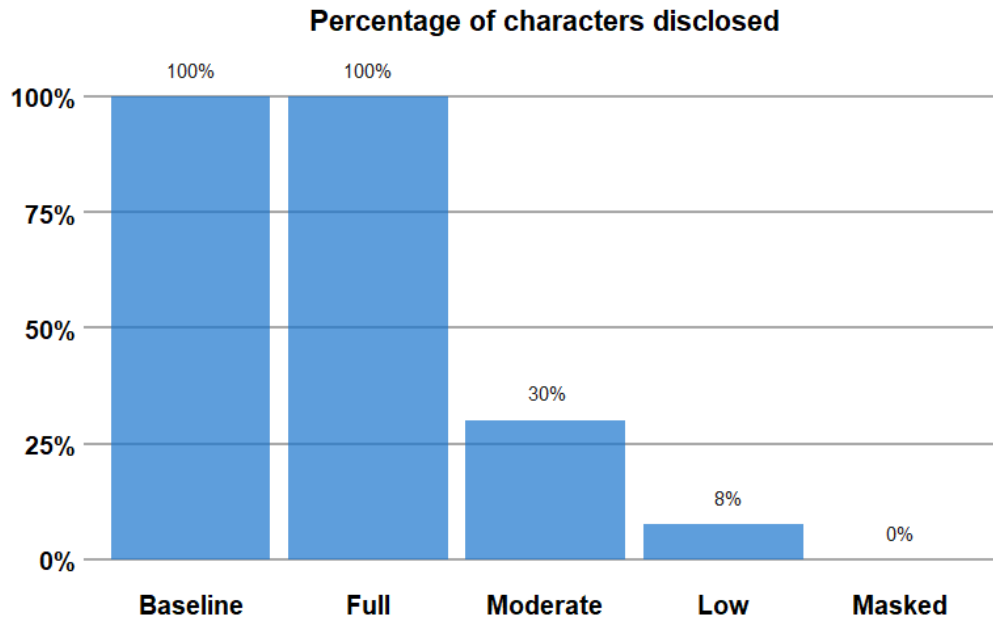
We summarize the differences among the conditions above and in Figure 9:

- **Baseline (full disclosure with no markup):** This condition displayed the full information from all records. As the baseline condition, no visual markup was available to highlight differences, and name frequency indicators were not included. This condition represents how record linkers would normally view records without any privacy protection, as it is similar to the conventional method used at most RL centers worldwide.
- **Full (full disclosure with markup):** This condition also displayed the full information from all records. No data values were hidden. In this view, pairs of records were augmented with graphical icons and color-coded text to highlight the differences between paired records, and frequency icons were included.

- **Moderate (moderate disclosure with markup):** The goal for this condition was to hide information except for the most relevant items believed to assist decision-making. Information was hidden in columns for pairs having the same values for both records, and check marks were instead shown to indicate matching values. Because identification (ID) numbers (such as SSNs) are often considered highly sensitive types of information and the raw value is not useful information for linkage decisions, full IDs were never revealed in this mode. Supplemental visual markup (difference icons, colored text, and frequency icons) was again used to highlight differences (the same as in *full*).
- **Low (low disclosure with markup):** The goal for this condition was to reveal as few data characters as possible while showing how pairs were different. As in the *moderate* condition, check marks were shown instead of values when the data in columns were the same, and visual markup (difference icons, colored text, and frequency icons) was again used to highlight any differences between 2 records. Unlike the *moderate* condition, little information was shown where there were differences. If a small number of characters in a field were different in 2 records, asterisks (*) were used to indicate matching characters, and only the values of the different characters were shown. For greater differences, no characters were shown, and the red *different* icon was shown. Gender was always visible in this mode to support decisions that required knowing the gender of the person without seeing the full name.
- **Masked (masked disclosure with markup):** This condition represents legally deidentified data, which shows no identifying data values and fully prioritizes privacy over information disclosure. Not a single actual character is revealed, and users must rely entirely on the supplemental visual markup (icons, colored symbols, and frequency icons). Check marks again denote matching columns. Representation of differing fields is most similar to the *low* condition, except the characters that are different are represented by different symbols (& and @) rather than their actual values, and values for gender are always hidden.

These conditions allowed us to test our hypotheses about the effects of different levels of information disclosure and the influence of supplemental markup. Figure 10 shows the average percentages of characters disclosed under the different conditions. The *baseline* and *full* conditions both show the values of all characters in the records, but the value hiding and character masking of the other conditions greatly reduce the amount of visible characters.

Figure 10. Differences in Percentage of Character Values Revealed With the Different Conditions Applied to the Generated Test Data Used for the Experiment



The different experimental conditions controlled the level of information disclosed to participants. This bar chart shows the differences in percentages of character values revealed with the different conditions applied to the generated test data used for the experiment. Percentages are relative to the number of characters in the baseline condition, which shows 100% disclosure of all characters. The moderate, low, and masked conditions hide matching characters and use character masks to greatly reduce the amount of visible characters. Reproduced from Ragan et al. 2018 CHI Conference on Human Factors in Computing Systems. <https://dl.acm.org/doi/10.1145/3173574.3173900>. Reprinted with permission from CHI'18 (Copyright ©2018). Association for Computing Machinery. All Rights Reserved.

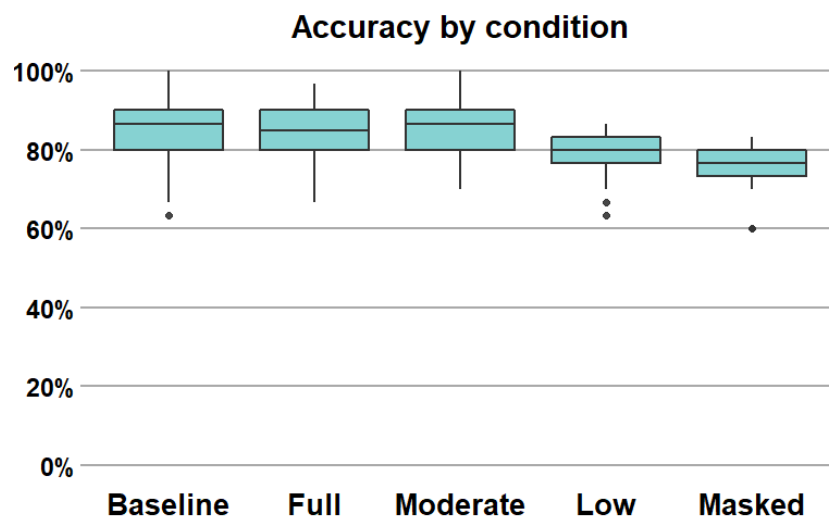
Results. This experiment (N = 104) focused on an evaluation of the quality of human decision-making with the visual data-masking technique. Table 2 presents the final summary characteristics of the participant samples, demonstrating reasonable balance. The accuracy of matching under the *low* and *masked* conditions was statistically significantly worse than that under the *full* and *moderate* conditions ($P < .05$). There was a near-significant difference ($P = .087$) between *low* and *masked* (Figure 9). These results support both H1 and H2, which indicate that participants who viewed only 30% of data details (*moderate* condition) had decision quality similar to that of those who had *full* (100%) access (Figure 11). H3 is concerned with differences between the baseline and the addition of supplemental markup, so we compared outcomes from the *baseline* and *full* conditions. No significant difference was found for accuracy ($P = .96$) (Figure 11) or time ($P = .58$). Thus, with no detected effects on time or accuracy, we reject H3

that the additional supplemental markup improved linkage performance. However, based on qualitative feedback, the use of supplemental markup was still important to maintain performance, as data elements were hidden under the other conditions. See Tables 3 through 8 for more details, including effect sizes.

Table 2. Summary of Participant Characteristics by Condition for Study A.1

	Measure	Baseline	Full	Moderate	Low	Masked	Total
n		20	20	23	21	20	104
Age, y	Mean	23.6	22.6	23.8	24.5	24.6	23.8
	SD	4.1	3.1	2.8	2.8	5.6	3.8
Male	No.	11	12	15	11	12	61
	%	55	60	65	52	60	59
Quantitative department	No.	12	15	18	16	14	75
	%	60	75	78	76	70	72
Graduate Student	No.	11	10	14	16	9	60
	%	55	50	61	76	45	58

Figure 11. Record Linkage Accuracy for the 5 Conditions



Reproduced from Ragan et al. 2018 CHI Conference on Human Factors in Computing Systems. <https://dl.acm.org/doi/10.1145/3173574.3173900>. Reprinted with permission from CHI'18 (Copyright ©2018). Association for Computing Machinery. All Rights Reserved.

Table 3. Accuracy of Linkage Decisions (% of Correct Responses) by Condition

Condition	n	Mean	SD
Baseline	20	84.83	9.7
Full	20	84.17	7.64
Low	21	78.1	5.83
Masked	20	74.5	7.2
Moderate	23	84.49	7.433

Table 4. Completion Time by Condition

Condition	n	Mean, min	SD, min
Baseline	20	9.3	2.56
Full	20	10.66	5.79
Low	21	11.33	6.29
Masked	20	11.02	3.56
Moderate	23	12.79	7.86

Table 5. Participant Confidence in Linkage Decisions by Condition (Scale, 1-3)

Condition	n	Mean	SD
Baseline	20	2.6	0.32
Full	20	2.29	0.33
Low	21	2.32	0.37
Masked	20	2.22	0.38
Moderate	23	2.44	0.35

Table 6. Accuracy Effect Size

Condition 1	Condition 2	Cohen <i>d</i>	Cohen <i>d</i> low	Cohen <i>d</i> high	<i>t</i>	<i>P</i> value
Baseline	Full	0.07	-0.55	0.7	0.24	.81
Baseline	Low	0.83	0.22	1.54	2.68	.01
Baseline	Masked	1.186	0.57	1.94	3.83	.01
Baseline	Moderate	0.04	-0.59	0.67	0.13	.90
Full	Low	0.88	0.27	1.59	2.85	.01
Full	Masked	1.28	0.65	2.05	4.12	.01
Full	Moderate	0.04	-0.58	0.67	-0.14	.89
Low	Masked	0.54	-0.07	1.21	1.75	.09
Low	Moderate	0.94	0.33	1.65	-3.19	.01
Masked	Moderate	1.34	0.72	2.11	-4.47	.01
Baseline	Full	0.07	-0.55	0.71	0.24	.81

Table 7. Time Effect Size

Condition 1	Condition 2	Cohen <i>d</i>	Cohen <i>d</i> low	Cohen <i>d</i> high	<i>t</i>	<i>P</i> value
Baseline	Full	-0.3	-0.95	0.32	-0.96	.35
Baseline	Low	-0.41	-1.07	0.2	-1.37	.18
Baseline	Masked	-0.54	-1.22	0.07	-1.76	.09
Baseline	Moderate	-0.57	-1.24	0.04	-2.01	.05
Full	Low	-0.11	-0.74	0.51	-0.36	.72
Full	Masked	-0.07	-0.71	0.55	-0.24	.81
Full	Moderate	-0.3	-0.95	0.31	-1.02	.31
Low	Masked	-0.06	-0.69	0.57	0.2	.84
Low	Moderate	-0.2	-0.84	0.42	-0.68	.50
Masked	Moderate	-0.28	-0.92	0.34	-0.97	.34
Baseline	Full	-0.3	-0.95	0.32	-0.96	.35

Table 8. Confidence Effect Size

Condition 1	Condition 2	Cohen <i>d</i>	Cohen <i>d</i> low	Cohen <i>d</i> high	<i>t</i>	<i>P</i> value
Baseline	Full	0.92	0.3	1.63	2.95	.01
Baseline	Low	0.78	0.17	1.47	2.54	.02
Baseline	Masked	1.05	0.44	1.79	3.4	.01
Baseline	Moderate	0.45	-0.16	1.11	1.52	.14
Full	Low	0.08	-0.54	0.72	-0.27	.79
Full	Masked	0.2	-0.42	0.84	0.65	.52
Full	Moderate	0.43	-0.19	1.08	-1.43	.16
Low	Masked	0.27	-0.35	0.92	0.88	.38
Low	Moderate	0.32	-0.29	0.97	-1.09	.28
Masked	Moderate	0.6	-0.01	1.27	-1.98	.05
Baseline	Full	0.92	0.3	1.63	2.95	.01

These results demonstrate that it is possible to greatly limit the amount of identifying information available to human decision makers without negatively affecting human decision-making. However, the findings also show there is a limit to how much data can be hidden before negatively influencing the quality of judgment in RL decisions involving person-level data. Despite the reduced accuracy with extreme data hiding, the study demonstrates that with proper interface designs, many correct decisions can be made with even legally deidentified data that are fully masked (as seen in Table 3, there was 74.5% accuracy with fully masked data, compared with 84.1% with full access). Thus, when legal requirements only allow for deidentified data access, the use of a well-designed interface can significantly improve data utility.

Study A.2: Dynamic Design – Interactive On-Demand Incremental Information Disclosure

Using the well-designed visual masks in study A.1, we expanded the interface and conducted a controlled experiment to evaluate how different mechanisms for privacy protection affect information access and decision-making for RL tasks.

Hypotheses. Our overarching goal is to design and evaluate effective ways to discourage unnecessary information disclosure without increasing linkage errors. In this experiment, we test the effect of the following 3 mechanisms: (1) an interactive, clickable on-demand disclosure interface; (2) transparent accountability through measuring the real-time risk on a meter; and (3) enforcing limitations on disclosures through a prespecified budget on the meter. Our evaluation of these mechanisms follows 3 respective hypotheses:

- **H1:** We hypothesize that an appropriate on-demand and incremental disclosure interface can significantly reduce disclosure without compromising decision quality. This is the main premise behind our design for interactive, on-demand information access. An explicit click by the user is required to disclose any piece of PII, which means that all clicks, and thus disclosures, can be tracked. Given that users will have the ability to look at any part of the PII, there should be no impact on the quality of the decision.
- **H2:** The second hypothesis is that the addition of the feedback mechanism, which quantifies and provides a real-time display of consequences of the click, can better inform the decision to access information, and hence encourage only the most-needed disclosure. Quantification of the risk and visibility of this information for all relevant parties (eg, users, managers, compliance) will discourage misuse of PII and encourage accountable use of PII through transparency.
- **H3:** The third hypothesis is that when providing feedback on disclosure, enforcing a limit on privacy disclosure through a prespecified budget will change disclosing behavior to tend toward the given limit. That is, we expect people will naturally try to use the full available budget. In other words, if the limit is set high, higher levels of disclosure will occur (H3.1). On the other hand, if the limit is set too low, disclosure levels will be forced to be lower, but decision quality will be negatively affected (H3.2). Hypothesis H3.2 follows the results from study A.1, which provided evidence of a limit to how much data can be hidden before negatively influencing the quality of judgment in decisions involving person-level data.

Experiment design. To address our hypotheses, the experiment followed a between-subjects design with the following 5 conditions:
















- **Fully open:** Nonclickable interface with all details already visible. This was the baseline condition used to study the effect of different mechanisms. It used the static full-

disclosure interface with visual discrepancy highlighting and frequency metadata, but no data were hidden.

- **No meter:** Clickable on-demand disclosure with no feedback meter and no limit. The goal for this condition was to test the effect of using an interactive on-demand interface on the amount of disclosure and decision quality. The initial interface starts with a fully masked display with markups, and users can click to disclose more information. The KAPR feedback meter was not shown, and there was no limit to information access.
- **Unlimited meter:** Clickable on-demand disclosure with an unlimited feedback meter. The goal of this condition was to test the effect of adding the KAPR meter (see top of Figure 6, marked 2 in red) to display the potential real-time increase in risk for any given disclosure to inform the decision to view the data. There was no limit to disclosure in this condition.
- **High limit:** Clickable on-demand disclosure with a feedback meter and a high limit. This condition tests the effect of enforcing a prespecified limit on the privacy budget (see top of Figure 6, marked 3 in red) indicated by a thick red line on the meter. This condition sets a moderate disclosure limit believed to be sufficient to make good linkage decisions. The specific limit under this condition was a 35.7% to 37.8% KAPR score, depending on the specific data set. This amount was chosen based on the *moderate*-level from study A.1. The prior study found this level of disclosure had comparable decisions as full disclosure, so we would expect good linkage performance if participants used the full budget.
- **Low limit:** Clickable on-demand disclosure with a feedback meter and a low limit. This condition is similar to the previous condition in enforcing a limit on the privacy budget (see top of Figure 7, marked 3 in red). This condition sets a lower limit with KAPR scores ranging from 5.02% to 6.48%, depending on the data set. This level was again chosen based on study A.1, which found reductions in linkage decisions with this amount of static disclosure. In the current study, users choose which details to access interactively, as needed. Thus, this condition tests whether total disclosure levels can come down to these low levels without compromising linkage decisions when interactive disclosure is used.

Figure 12 shows a simplified summary of the differences among the 5 conditions. The conditions allowed us to test our hypotheses about the effects of different mechanisms to discourage unnecessary disclosure.

Figure 12. Visual Summary Representing the Differences of the 5 Experimental Conditions In the Evaluation

Condition	Default Masking	On-demand Interface	Meter & Limit
Fully open			
No meter			
Unlimited meter			
High Limit			
Low limit			

Reproduced from Kum et al. Fifteenth USENIX Conference on Usable Privacy and Security.

<https://dl.acm.org/doi/10.5555/3361476.3361489>. Reprinted with permission from SOUPS'19 (Copyright ©2019).

All Rights Reserved.

Results. The experiment (N = 120) studied RL behavior with a focus on the just-in-time or on-demand interactive interface design for incrementally disclosing partial information only when needed. Table 9 presents the final summary characteristics of the samples demonstrating reasonable balance. We evaluated the approach with a controlled experiment of how different types of feedback and access restrictions affect human decision-making quality, speed, and access behavior.

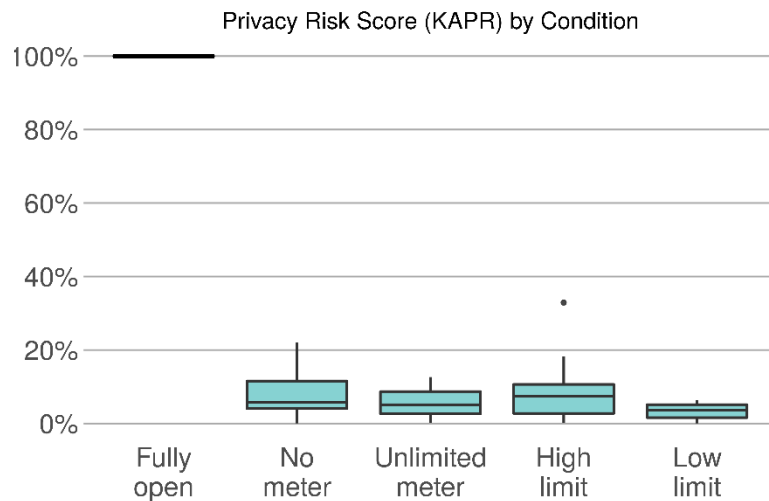
Table 9. Summary Characteristics by Condition for Study A.2

	Measure	Fully open	No meter	Unlimited meter	High limit	Low limit	Total
n		22	23	26	26	23	120
Age, y	Mean	24.3	22.7	22.4	24.2	24.6	23.6
	SD	3.4	4.1	4.9	3.7	5.9	4.5
Male	No.	11	7	11	13	11	53
	%	50	30	42	50	48	44
Quantitative department	No.	12	10	13	15	12	62
	%	55	43	50	58	52	52
Graduate student	No.	14	11	11	15	12	63
	%	64	48	42	58	52	53

The results provided support for H1 when we compared *fully open* with the *no-meter* condition to determine how much more we can reduce disclosure using the interactive interface. Even with very low levels of disclosure in *no meter* (only 7.85%, compared with 100% in *fully open*; Figure 13), the error rate did not increase significantly compared with *fully open* (Figure 14). A student *t* test did not find a significant difference between the error rates ($P = .22$). Though no difference was found, we cannot definitively claim that the on-demand disclosure method did not induce an increase in the error rate.

To address H2, we compared *no meter* and *unlimited meter* to determine if a feedback meter is effective in reducing unnecessary disclosure. The study results did not provide evidence for H2. The quality of decision and completion time were similar (Figures 14 and 15), and although adding the feedback meter to the interactive on-demand disclosure reduced the KAPR score from 7.85% to 5.33% (Figure 13), this difference was not statistically significant. However, the relatively low *P* value ($P = .07$) suggests that the results may be inconclusive, and it motivates further study, especially considering other findings indicating that people may change privacy behavior with appropriate feedback, which is consistent with the literature.^{36,37}

Figure 13. KAPR Privacy Scores for the 5 Conditions



Abbreviation: KAPR, k-Anonymity Privacy Risk.

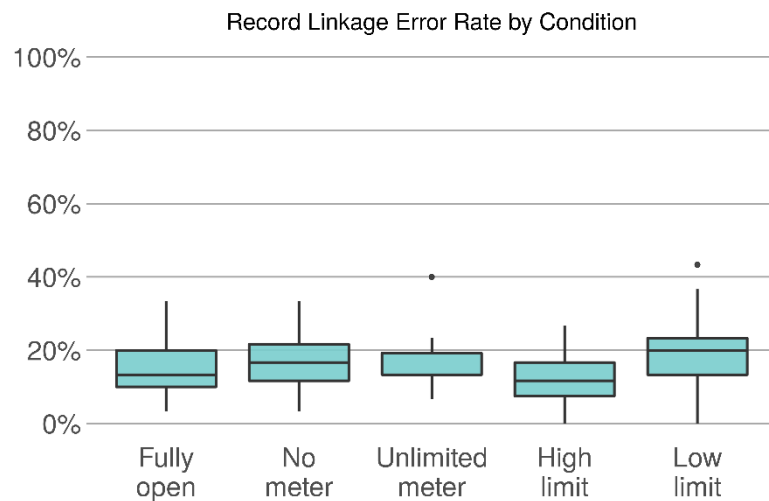
Lower scores indicate lower risk. Note that the fully open condition has 100% privacy risk score due to all characters being visible by default.

Reproduced from Kum et al. Fifteenth USENIX Conference on Usable Privacy and Security.

<https://dl.acm.org/doi/10.5555/3361476.3361489>. Reprinted with permission from SOUPS'19 (Copyright ©2019).

All Rights Reserved.

Figure 14. Percentage of Incorrectly Linked Pairs From the 5 Conditions



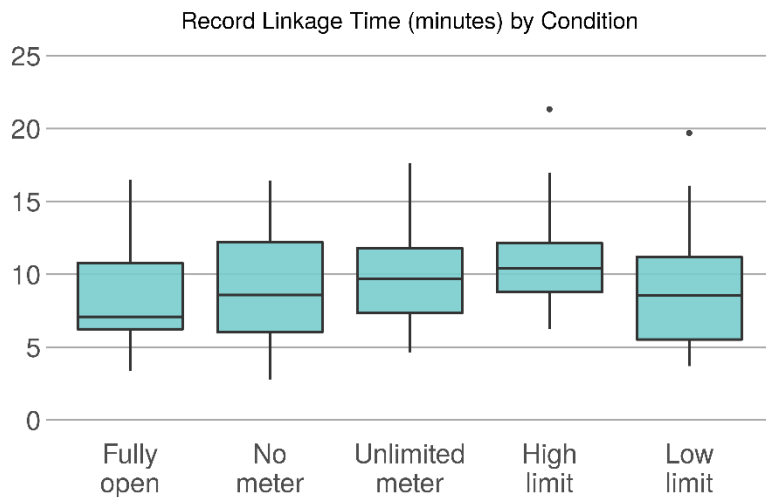
Lower values indicate better performance. The median is the center horizontal line of data. The interquartile range box represents the middle 50% of the data. The whiskers, extending from either side of the box, represent the ranges for the bottom 25% and the top 25% of the data values, excluding outliers (dots).

Reproduced from Kum et al. Fifteenth USENIX Conference on Usable Privacy and Security.

<https://dl.acm.org/doi/10.5555/3361476.3361489>. Reprinted with permission from SOUPS'19 (Copyright ©2019).

All Rights Reserved.

Figure 15. Time Taken to Complete the Linkage Task for the 5 Conditions



Reproduced from Kum et al. Fifteenth USENIX Conference on Usable Privacy and Security. <https://dl.acm.org/doi/10.5555/3361476.3361489>. Reprinted with permission from SOUPS'19 (Copyright ©2019). All Rights Reserved.

Finally, H3 compared the *unlimited meter*, *high limit*, and *low limit* to evaluate the impact of different levels of limit on the amount of disclosure and quality of RL. H3.1 did not hold in the comparison between *unlimited meter* and *high limit*. There were no statistical differences in error rate, KAPR score, or the time between these conditions. Although disclosure levels were higher under the high-limit condition than for not having a limit (7.87% vs 5.33%, respectively), the amount of expended disclosure in the unlimited condition (mean [SD], 36.7% [0.81%]) did not near the given budget limit. On average (SD), participants used only 21.4% (19.1%) of the given budget. Although a high limit did nudge participants to disclose slightly more, the study found that participants were still careful when disclosing the data. We believe this is the result of the short tutorial which emphasized opening only what was needed and participants being privacy conscious.

However, H3.2 did hold in the comparison between *high limit* and *low limit*. For participants given the *low-limit* condition, the KAPR score was less than half (mean [SD], 3.22% [2.12%]) of those given the *high-limit* condition (mean [SD], 7.87% [7.09%]), with evidence of differences in the risk scores ($P < .001$). There were also significant differences in the error rate scores between the modes ($P = .012$). The error results indicate that the quality of human decisions will suffer if low disclosure limits are enforced.

See Tables 10 through 15 for more details including effect sizes. In sum, the on-demand interactive interface reduced privacy risk to only 7.85% (compared to 100% with all data disclosed), with little to no impact on decision quality or completion time. The results serve as evidence that the incremental disclosure method can be highly effective for ensuring legal compliance with the “minimum necessary” and accountable access requirements.

Table 10. Error Rate of Linkage Decisions (% of Wrong Responses) by Condition

Condition	n	Mean	SD
Fully open	22	15.35	6.79
No meter	23	17.15	8.68
Unlimited meter	26	15.78	6.56
High limit	26	12.82	7.04
Low limit	23	19.16	9.81

Table 11. Record Linkage Time by Condition

Condition	n	Mean, min	SD, min
Fully open	22	8.57	3.48
No meter	23	8.82	3.45
Unlimited meter	26	10.07	3.65
High limit	26	11.03	3.32
Low limit	23	9.04	4.12

Table 12. Privacy Risk Score (KAPR) by Condition

Condition	n	Mean	SD
Fully open	22	0	0
No meter	23	7.85	5.23
Unlimited meter	26	5.33	3.79
High limit	26	7.87	7.09
Low limit	23	3.22	2.12

Abbreviation: KAPR, k-Anonymity Privacy Risk.

Table 13. Error Effect Size

Condition 1	Condition 2	Cohen <i>d</i>	Cohen <i>d</i> low	Cohen <i>d</i> high	<i>t</i>	<i>P</i> value
Fully open	No meter	-0.23	-0.82	0.34	-0.78	.44
Fully open	Unlimited meter	-0.06	-0.65	0.51	-0.23	.82
Fully open	High limit	-0.36	-0.96	0.21	1.26	.21
Fully open	Low limit	-0.44	-1.05	0.12	-1.52	.14
No meter	Unlimited meter	-0.18	-0.76	0.39	0.61	.54
No meter	High limit	-0.54	-1.15	0.02	1.9	.06
No meter	Low limit	-0.21	-0.8	0.36	-0.74	.47
Unlimited meter	High limit	-0.43	-1.03	0.13	1.57	.12
Unlimited meter	Low limit	-0.40	-1	0.16	-1.4	.17
High limit	Low limit	-0.74	-1.37	-0.17	-2.57	.10
Fully open	No meter	-0.23	-0.82	0.34	-0.78	.44

Table 14. Time Effect Size

Condition 1	Condition 2	Cohen <i>d</i>	Cohen <i>d</i> low	Cohen <i>d</i> high	<i>t</i>	<i>P</i> value
Fully open	No meter	-0.07	-0.65	0.51	-0.23	.82
Fully open	Unlimited meter	-0.41	-1.02	0.15	-1.45	.15
Fully open	High limit	-0.71	-1.34	-0.15	-2.49	.02
Fully open	Low limit	-0.12	-0.71	0.45	-0.42	.68
No meter	Unlimited meter	-0.35	-0.94	0.22	-1.24	.22
No meter	High limit	-0.64	-1.27	-0.08	-2.28	.03
No meter	Low limit	-0.06	-0.64	0.52	-0.2	.84
Unlimited meter	High limit	-0.27	-0.86	0.29	-1	.32
Unlimited meter	Low limit	-0.26	-0.85	0.31	0.92	.36
High limit	Low limit	-0.53	-1.14	0.04	1.84	.07
Fully open	No meter	-0.07	-0.65	0.51	-0.23	.82

Table 15. KAPR Effect Size

Condition 1	Condition 2	Cohen <i>d</i>	Cohen <i>d</i> low	Cohen <i>d</i> high	<i>t</i>	<i>P</i> value
Fully open	No meter	-2.04	-2.87	-1.41	-7.21	<i>P</i> <0.01
Fully open	Unlimited meter	-1.88	-2.67	-1.26	-7.18	<i>P</i> <0.01
Fully open	High limit	-1.48	-2.21	-0.89	-5.66	<i>P</i> <0.01
Fully open	Low limit	-2.07	-2.9	-1.44	-7.3	<i>P</i> <0.01
No meter	Unlimited meter	-0.55	-1.16	0.01	1.91	.06
No meter	High limit	0	-0.58	0.57	-0.01	.99
No meter	Low limit	-1.14	-1.82	-0.57	3.94	<i>P</i> <0.01
Unlimited meter	High limit	-0.44	-1.04	0.12	-1.61	.12
Unlimited meter	Low limit	-0.67	-1.29	-0.11	2.44	.02
High limit	Low limit	-0.86	-1.5	-0.3	3.18	<i>P</i> <0.01
Fully open	No meter	-2.04	-2.87	-1.41	-7.21	<i>P</i> <0.01

Abbreviation: KAPR, k-Anonymity Privacy Risk.

Study B: Summative Evaluation

Objective

The goal of the summative evaluation was to study the use of MiNDFIRL in real settings with data workers tasked with linking real data. The formative evaluations (ie, studies A.1 and A.2 described above) provided empirical data for technique verification and established a foundational knowledge for trade-offs among decision-making quality, data privacy, and access behaviors using interactive on-demand techniques with a privacy budget to limit total access. However, the formative studies were conducted with the software configured with only small data sets for testing purposes. Through this summative evaluation (study B), we sought to identify new concerns and recommendations for consideration when moving the approach from the research stage to full application development. In study B, we aimed to investigate whether the findings from the prior studies (ie, studies A.1 and A.2) would be observed in the new, more realistic, and more complex operational scenarios. Specific issues of interest included interface design appropriateness, effects of the on-demand disclosure technique on information access decisions, and adaptability of MiNDFIRL to different forms of data while fitting within a complete end-to-end data pipeline for data cleaning and linkage. While the prior formative evaluations heavily emphasized quantitative measures, the summative evaluation prioritizes qualitative data from participant experiences and feedback to provide a more holistic understanding of implications and feasibility of the research techniques. Given the desired linkage specialization for the target usage context, the study B evaluation takes the form of case studies with small groups of data workers.

Methods

Study design. The study included 2 case studies at 2 locations, (i) UTH and (ii) UAB. Both cases involved RL with data sets sampled from the corresponding locations. The studies consisted of linkage projects with teams involving 4 data reviewers with 1 of the reviewers also playing a second role as team manager. The data reviewers were responsible for using MiNDFIRL to review data discrepancies to perform RL. The manager was responsible for

configuring the software for allowable data fields, setting the privacy budget for data access, and assigning data pairs to the data workers. The process of setting up the linkage projects and configuring MiNDFIRL served as a proof of concept for applying the software techniques and integrating in the respective real data environment with consideration of specific data needs and properties (eg, which data fields, how many fields, which data to link, and coordination between human RL processes and automated RL methods).

The software included a tutorial explaining the interface and RL task. The data reviewers were tasked with reviewing the assigned sets of discrepancy pairs in the software. For each pair, participants were required to indicate whether the 2 entities should be considered the same or different entities, and each decision also includes a level of confidence (low, medium, or high) for the linkage decision (see Figure 4).

After the research team and team managers configured the data projects and assigned linkage sets to the team members, the data reviewers used the software to complete their linkage assignments in 2 separate sessions over a period of 1 week. Session times varied, with linkage sessions taking approximately 60 to 90 minutes for the first session (Figure 3, reviewer1 and reviewer2) and 30 to 40 minutes for the second session (Figure 3, reviewer3 and reviewer4 where needed). Teams conducted the linkage activities independently and asynchronously at their respective locations with reminders and progress checks by the research team. Data reviewers were asked to complete brief experience questionnaires after each period of RL. The purpose of the questionnaires was to capture quick and lightweight notes of any issues, challenges, or thoughts immediately after using the software. Questionnaire prompts encouraged participants to record notes about general software usage or frustrations from each usage period.

Because the goal was to achieve the highest possible quality of data integrity through the RL process, the case study included an additional step to resolve differences in the data review and decision-making processes. After all team members completed their linkage assignments, the software flagged pairs that were inconclusive among team members (ie, any time 2 reviewers indicated 2 data rows correspond to the same entity while the other 2

reviewers indicated they were different entities). Each team then participated in a “conflict-resolution” discussion (Figure 3, “Open discussion”) in which the team members all reviewed the disagreement cases together. To facilitate this discussion, the software provides a special viewing mode that allows team members to review these cases and see the responses of other team members. The manager led the teams in discussion of these special cases together (synchronously, via video conferencing) until the team finally decided on a consensus for all pairs. Using the software’s data-masking method, the software showed the pairs with the union of disclosed details among team members (in other words, any details for a particular pair that was disclosed by any team member would be visible during the conflict-resolution phase). The manager was able to disclose additional data details for each pair as needed during this phase. For rare cases where team consensus could not be reached with limited data values, the manager could reference the complete records for the discrepancy and make a final determination.

Following the conflict-resolution session, the team participated in a group discussion led by a member of the research team. The scope of the discussion included both RL and conflict resolution. The discussion followed the format of a semistructured interview to collect feedback about (a) general system usage and processes, (b) strategies and decision-making with the on-demand features, (c) understanding or challenges with the user interface, and (d) general recommendations, problems, and feedback. This data collection was conducted synchronously with the discussion format chosen to facilitate clarification and encourage discussion to aid a more complete level of understanding.

Participants. The case studies included a total of 12 data workers. The case study at UTH included 8 participants and integrated members of the research team with the data workers. The study consisted of 2 teams of 4 (1 manager and 3 data workers). The case study with the UAB included 4 participants working as a single team. Participant backgrounds and experience with data linkage varied.

Data configuration. Both case studies were conducted with data configured based on the location (UTH or UAB). The first case study at UTH used the “gold-standard” benchmark RL data set derived from UTH’s clinical data warehouse containing 2.61 million distinct medical record numbers (including potentially duplicate patient records)¹¹ on a Linux system. The benchmark data were developed from 10 million record pairs generated from the electronic health record (EHR) patient database by 6 reviewers who manually reviewed 20 000 randomly selected, potential match record-pairs to identify matches.¹¹ For the linkage activity, 8 fields were included: first name, middle name, last name, date of birth, SSN, gender, address, and phone number. Starting with the 20 000 labeled pairs, our study team used 10 000 labeled pairs to build an automated random-forest linkage model. We then used this model to study the full linkage process with the other 10 000 labeled pairs. After the automatic linkage, 303 uncertain pairs were selected for manual review based on low certainty by the algorithm.

The second case study at UAB used the rheumatic and arthritic patient data from ArthritisPower, a PPRN that was previously part of the National Patient-Centered Clinical Research Network (PCORnet), with 18 240 unique patient IDs on a Windows system. The fields used for the linkage case study included record ID, patient first name, patient last name, date of birth, sex, race, state, ZIP Code, email address, rheumatologist name, and rheumatologist National Provider Identifier. For the study, the data source was used to generate 1055 unique pairs based on records with identical matching on the following variables: (1) first name + last name, (2) first name + date of birth, and (3) last name + date of birth. Then, we used our random-forest–trained model from the UTH study for automatic RL on these pairs and adjusted the manual review thresholds to determine the 187 uncertain pairs that required manual review by people.

Results

Data flow. Figures 16(a) and (b) depicts the full data flow for the 2 site studies at UTH and UAB. Participants had no issues getting MiNDFIRL to run on both Linux and Windows systems as well as using it on both EHR and patient-generated data. Most of the manually

reviewed pairs were easily identified as a match or not by 2 independent data workers with no disagreement in the first review. The pairs with disagreement (ie, no match) were then reviewed by 2 more independent data workers (Figure 16). Three of the 4 reviewers agreed on most of these pairs, leaving only a small number of pairs, 19 pairs and 24 pairs, to be reviewed together at a meeting. Consensus was reached for all pairs except for 1 pair at UAB, which required a final determination by the UAB manager. In total, 232/620 (37%) and 84/623 (13%) more matches were found through the manual review process, but this required separating these matched pairs out from the full uncertain pair set, which were 77% (matched pairs/the full uncertain pair set = 232/303) and 45% (84/187) each. Since the random-forest model could not separate these out, it was important to use MiNDFIRL to manually find the additional matches without increasing the rate of false matches.

Figure 16 (a). Data Flow for UTH and UAB Study

UTH (EHR): KAPR \approx 36 % (Linux)

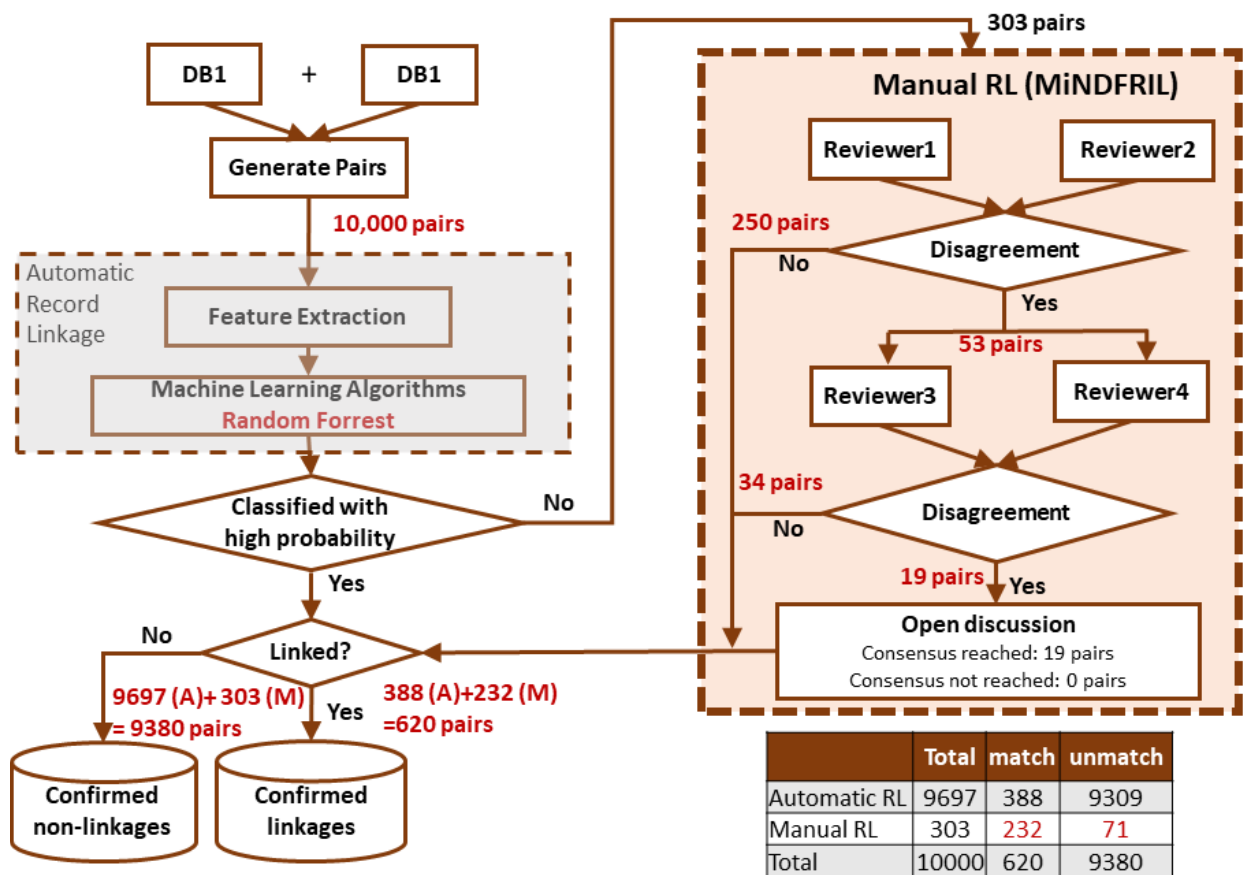
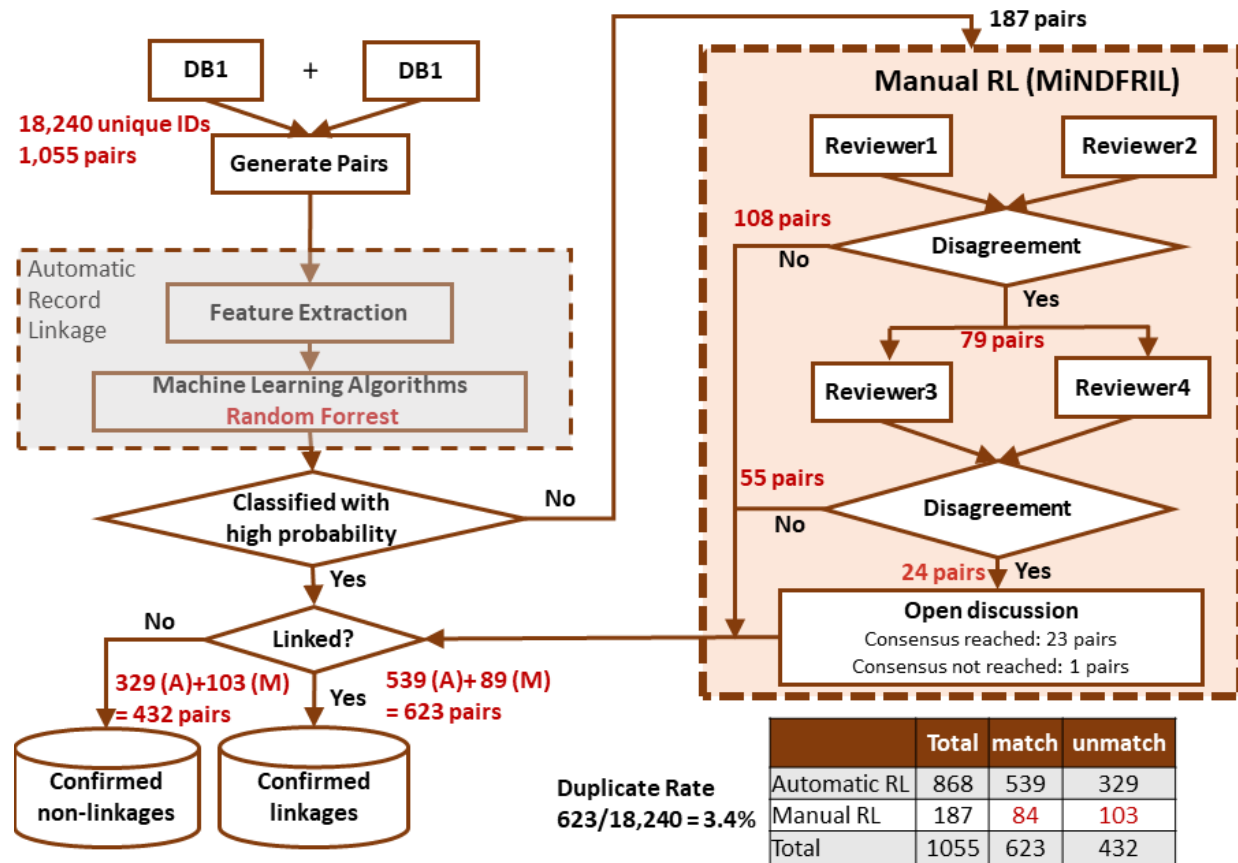


Figure 16(b). Data flow for UTH and UAB study (cont'd)

UAB (Patient Generated Data): KAPR \approx 30% (Windows)



Abbreviations: A, automatic; DB, data set; EHR, electronic health record; IDs, identifications; KAPR, k-Anonymity Privacy Risk; M, manual; MiNDFIRL, MInimum Necessary Disclosure For Interactive Record Linkage; RL, record linkage; UAB, University of Alabama at Birmingham Health System; UTH, University of Texas Health Science Center at Houston.

General themes from UTH and UAB study. The core technique studied through this research is the interactive method for on-demand disclosure of hidden data values to limit the total amount of sensitive data that had to be viewed by human reviewers. The study results demonstrate that the designed techniques are effective for this purpose. The prior controlled experiments (studies A.1. and A.2) provided evidence that the masking and on-demand access techniques are effective in significantly reducing the need to access identifying data. The case studies (study B) serve as a proof-of-concept demonstration that the same techniques are effective in more-realistic RL settings. Both UTH and UAB cases were able to achieve comfortable human linkage with the default disclosure budget for identifying information of

30%, which is based on our findings from the formative studies. In the UTH study, the first name, last name, and date of birth were disclosed the most during manual linkage, with most of the privacy budget being spent on looking at first names. In the UAB study, email, date of birth, last name, and first name were disclosed the most during manual linkage, with most of the budget being spent on looking at email addresses. Since the case studies were based on real data without a clear “ground truth” for comparison, they are unable to provide definitive results for the effects on accuracy of decision-making about matching 2 records. Participant feedback did not indicate notable problems or limitations that could not be addressed by using MiNDFIRL’s disclosure techniques, though data workers did sometimes express the need to refer to full source records for specific problem cases. Such behavior still aligns with the intended design of limited access to identifiable personal data to make a record match on an as-needed basis, but the software streamlined the process to help workers access the appropriate individual record(s) from the full appropriate source, as needed.

Comments from the case study agreed with and reinforced the findings of the interface design from the prior controlled experiments. Though the 2 case studies used different specific fields in their respective data sets, both study sites used the same general masking and highlighting methods as studied in the experiments. The feedback indicated that the visual highlighting of discrepancies and addition of icons were effective for helping data workers easily identify differences between entities in data pairs. The appropriateness of this visual design aspect is of crucial importance for allowing users to understand discrepancies by applying the data-masking methods to reduce the needs for data disclosure, and the collection of studies has provided strong evidence that most of the implemented methods were easily understood without the need for elaborate explanation for different types of discrepancies (eg, character insertions or deletions, character swaps, field value swaps, whole-value differences).

However, not all interface icons were equally useful. Some data workers noted varying levels of difficulty in making use of the icons for the provided *name frequency* metadata. These data provided the relative frequency of first and last names included in the source data set for each entity in a pair; sometimes, knowing the number of instances of an item can assist in

determining the uniqueness of the name. In an effort to reduce information complexity provided to data workers, the interface provided frequency level as 4 ordinal categories: unique occurrence, rare occurrence, common, and highly common. Whether the frequency data were given attention and how much they affected decision-making varied according to personal strategies and preferences. While the name frequency information itself was considered valuable and useful, the frequency icons were sometimes challenging to interpret meaningfully in actual use. We suspect the best choices for specific distinctions for levels of frequencies will likely depend on the specific needs for a given project; the software may be more useful if the manager can customize how name frequency feedback is provided. Different icons may be needed for new levels, and in some cases, interval-level frequency information may be desired rather than ordinal categories.

As expected, the studies found that different data workers adopted different strategies and mindsets when conducting data linkage. For example, certain workers might give more attention to an *ID* field, while others might put more weight on a *date of birth* field for making linkage decisions. While not a problem, this finding does reinforce the importance of software that supports collaborative decision-making and conflict resolution to address between-worker differences and perspectives throughout the linkage process. Future iterations of software that supports our method might explore the integration of algorithmic techniques that can help log the history of data-access preferences by individual data workers (eg, allow worker A to see that worker B tends to reveal ID information for 70% of all access requests) to help facilitate a shared understanding of different perspectives and priorities during conflict-resolution discussions.

Different data workers also took different strategies for making use of the allowable disclosure limit or “privacy budget” for revealing data details. For instance, some adopted a more aggressive approach in opening more details early on despite the risk of exhausting the available budget, while others opted a more conservative approach of avoiding disclosure for the entire data set despite having a full budget available. The design rationale for budgeting on-demand disclosure is based on the assumption that data workers will only access more data

details when necessary for improved decisions. It is important to note that users should be encouraged to review more data details when they feel it will add value and improve linkage. The presence of strong differences in strategies might suggest that (a) for the case of aggressive disclosure, the available budget was too low for users to be confident in their linkage decisions; or (b) for the conservative strategist, the participant either did not seek to optimize decision quality or did not perceive a benefit to disclosing more details. Variation in strategy is common when freedom of choice and human decision-making are involved, though we expect that variance might be reduced through explicit instruction for recommended strategies and longer periods of practice to develop a practical sense of an optimal “spending” rate. Further investigation of strategies and budget usage over longer periods would be needed to appropriately adjust the budget in the software, and the existing support for budget adjustment would make this possible with the current software framework.

Aim 3: Develop 3 Companion Documents Through Participatory Action Research

The objective of aim 3 was to collaboratively create 3 companion template documents with patients and stakeholders: (1) a privacy statement, (2) an IRB application, and (3) a DUA. These documents are intended to improve communication and transparency between stakeholders (eg, patients, IRBs, and legal compliance staff) of secondary database research projects using MiNDFIRL to enhance privacy during RL.

Five studies informed the creation of these 3 documents. Those 5 studies were, in the order conducted, (1) ELSI nominal group technique (NGT) (study D.1); (2) patient NGT (study C.1); ELSI Delphi (study D.2); patient Delphi (study C.2); and a final frequently asked question (FAQ) evaluation survey (study C.3). The findings from each study informed subsequent research, and feedback was also iteratively incorporated into the software design. All studies were approved by the IRB at Texas A&M University.

Aim 3.1: Privacy Statement – FAQ Website for MiNDFIRL

The purpose of the aim 3.1 studies is to understand how best to communicate to the public complex issues relating to secondary database research. The aim 3.1 studies (C.1, C.2,

and C.3) collectively contribute to the collaborative creation of an FAQ to answer questions data subjects might have relating to the use of their data in research. In contrast to a privacy statement which forces patients to find answers that can be long, technical, and difficult to understand, FAQs present common questions and provide direct answers. The FAQ developed from the aim 3.1 findings provides answers to general questions and answers relating to secondary database research and is designed to foster an understanding of RL and the MiNDFIRL software.

Studies C.1, C.2, and C.3 aim to better understand patients' preferences and concerns related to the use of their data in secondary database research and to improve communication practices between researchers, health care entities, and patients. Communication is challenging in secondary database research because direct contact between data subjects and researchers is exceptionally rare (ie, because informed consent is commonly waived). Nevertheless, communication between patients and researchers is critical to facilitate trust and transparency and to integrate the patient voice in CER.

Studies C.1 and C.2 used NGT and Delphi methods with patient participants. These studies were informed by studies D.1 and D.2 (below), which solicited feedback from ELSI experts (eg, identified issues were incorporated in the initial draft FAQ). Study C.3 was an online survey to evaluate and solicit feedback on the final FAQ list developed from a more nationally representative sample.

Below we describe the objectives, methods, and results of these 3 studies. The material presented in this section previously appeared in the following 2 peer-reviewed publications:

- Study C.1 was published in Giannouchos T, Ferdinand AO, Ilangovan G, et al. 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. *J Am Med Inform Assoc.* 2021;28(8):1746-1755.³⁸
- Study C.2 was published in Schmit C, Ajayi KV, 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(12):e20783.
doi:10.2196/20783.³⁹

Study C.1 – patient NGT.

Objective. The aim of study C.1 was to qualitatively assess patients’ perceptions on the (1) benefits and (2) risks of using privacy-enhancing RL software, and (3) the additional information that patients would like if their medical data were used for research.

Methods

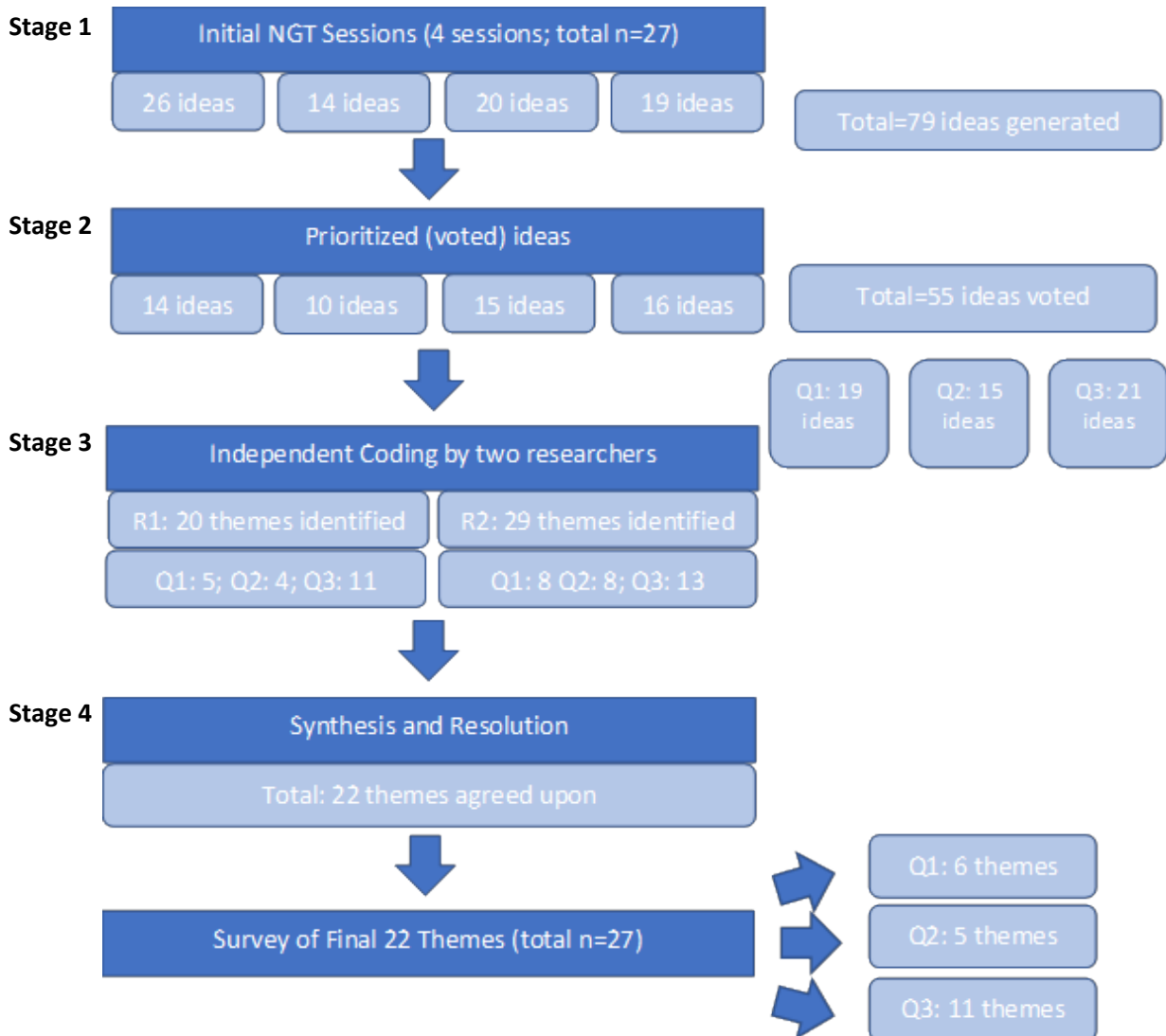
Design. We conducted 4 separate NGT sessions with different participants, each lasting approximately 120 minutes. The NGT method is 1 of the most commonly used qualitative methods for group decision-making processes to enable a group of people to generate and rank ideas on a topic that can be prioritized through discussion.⁴⁰⁻⁴² The NGT uses an ordered and collaborative approach to obtain reliable qualitative data.⁴¹ Used correctly, it is a useful and valid tool that can help identify and clarify problems.⁴¹ Accordingly, the NGT method was appropriate to identify key issues to address in a future FAQ document.

Our study groups ranged from 6 to 7 participants, which is within the recommended optimal range of 5 to 12 for NGT studies.⁴² A subsequent combined online survey was conducted to consolidate identified themes from 4 individual sessions to identify consensus priorities (Figure 17).

Participants and recruitment. Eligibility criteria for participants were English-speaking individuals at least 18 years of age with a chronic condition who had more than 3 health care provider visits within the previous year. We recruited patients using email lists from PPRNs, (ArthritisPower, COPD PPRN, Health eHeart, Interactive Autism Network, Mood Network, and PRIDENet) and employees and staff of a large university. The group of individuals who expressed interest in participating overrepresented women and White individuals. To improve representation, we randomly selected participants based on a stratified sample, oversampling male and non-White participants among those who expressed interest in participating.

Participants received a \$40 gift card and were entered a raffle to receive an additional \$100 gift card after completing the final online survey.

Figure 17. Idea Generation and Building Consensus Through 4 NGT Sessions and an Online Survey



Abbreviations: NGT, nominal group technique; Q, question; R, round.
 Reproduced from Giannouchos et al. *J Am Med Inform Assoc.* 2021;28(8):1746-1755. *J Am Med Inform Assoc* (Copyright ©2021). Journal of the American Medical Informatics Association. All Rights Reserved.

Individual NGT sessions. Three sessions were conducted via an online platform to allow nationwide participation. One on-site session was conducted in a computer lab. One researcher

with legal, privacy, and bioethics expertise led all NGT sessions. We finalized 3 questions for the NGT sessions after AC input and pilot testing to improve validity:

- Are there things you like about the software that you would tell your neighbors?
- Are there concerning things about the software that you would tell your neighbors?
- What more would you like to know?

At the start of each NGT session, participants completed a 20-minute online tutorial on privacy and RL. The tutorial included hands-on experience using MiNDFIRL for RL. Participants were then given access to an online document where they could enter their responses to the 3 questions. Participants were given 10 minutes per question to independently generate and record their ideas. Afterward, the facilitator led a discussion for each question, seeking clarification as appropriate and combining common ideas into themes. Participants were then asked to vote on the 2 most important themes per question.

Combining themes across all sessions. After all NGT sessions were completed, we deployed a survey to all prior NGT participants to identify the highest-priority issues among those identified in the smaller NGT groups. Two researchers independently conducted thematic analyses on the results to identify common themes across all sessions, excluding themes that obtained no votes in the individual sessions. The final list of themes was then created based on consensus between the researchers (Figure 17). The list was validated by 2 different researchers who mapped the final themes back to the ideas from each individual session. All participants were asked to rank the benefits and risk themes based on significance in a short online survey. For the third question (additional information needed), participants used a 5-point Likert scale for each theme to indicate the level of necessity to include the information in an FAQ for a research project using the software.

Results. In total, 27 patients participated in the 4 NGT sessions, and all 27 participated in the final voting and ranking. On average (SD), participants had a chronic condition for around 14 (13.7) years and around 5 (2.4) visits to their physician during the previous year, with a mean (SD) age of 48 (15.5) years (Table 16).

Table 16. Sociodemographic and Clinical Characteristics of Participants

	N=27
Average years with chronic condition(s) (standard deviation)	11.2 (SD 13.7)
Years with chronic condition(s)	
5 or less	30%
6 to 10	26%
11 to 15	15%
16 or more	30%
Average number of physician visits (standard deviation)	4.7 (2.4)
Top chronic conditions	
COPD	19%
Mental Health	19%
High blood pressure	11%
High cholesterol	7%
Irritable bowel syndrome (IBS)	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	
Medicare	15%
Medicaid	4%
Dual (Medicare & Medicaid)	11%
Private	59%
VA or DoD	4%
Other	7%
Average age (standard deviation)	48 (15.5)

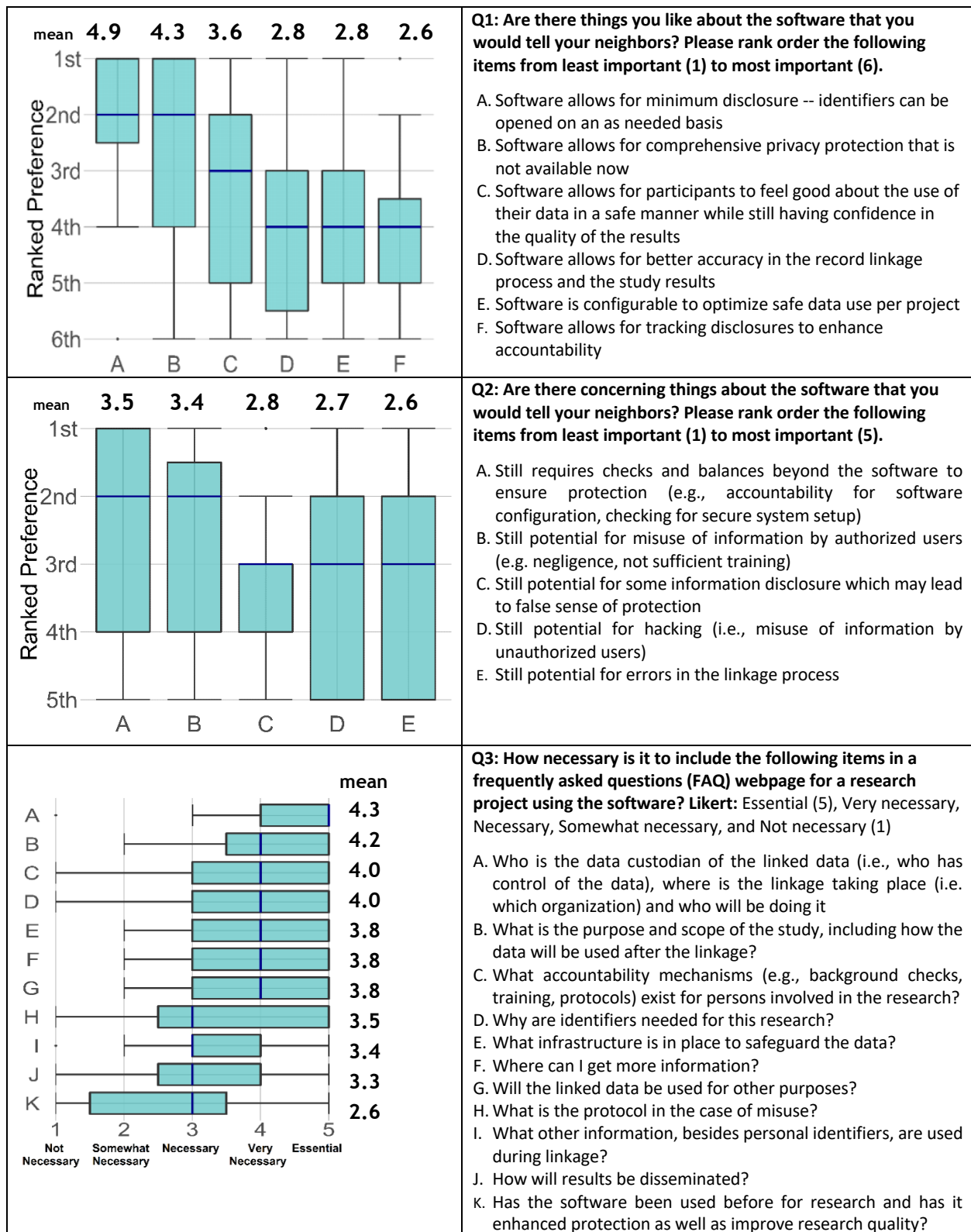
Age groups	
19-44	33%
45-64	48%
>65	19%
Gender	
Male	37%
Female	63%
Race/Ethnicity	
Non-Hispanic white	74%
Non-Hispanic black	7%
Hispanic	7%
Non-Hispanic Asian/Pacific Islander	11%
Income groups	
Less than \$25,000	15%
\$25,000-\$75,000	59%
\$75,000-\$125,000	15%
More than \$125,000	11%
Education level	
High school graduate or equivalent (GED)	7%
Some College	11%
College Graduate	56%
More than College	26%

Abbreviations: COPD, chronic obstructive pulmonary disease; DoD, Department of Defense; IBS, irritable bowel syndrome; VA, Veterans' Administration.

Reproduced from Giannouchos et al. *J Am Med Inform Assoc.* 2021;28(8):1746-1755. *J Am Med Inform Assoc* (Copyright ©2021). Journal of the American Medical Informatics Association. All Rights Reserved.

Across all 4 sessions, participants generated a total of 79 ideas for all 3 questions. Of those, 14 ideas did not receive any votes, leaving 55 ideas for thematic analysis. There were similarities, overlaps, and general consensus among most ideas and issues raised on each question across the 4 groups, which was an indicator of saturation. For example, “minimum disclosure” as a benefit came up in all 4 sessions. This process resulted in 22 themes in total, with 6, 5, and 11 themes for questions 1, 2, and 3, respectively (Figure 17).

Figure 18. Benefits, Risks, and Additional Information From the Final Online Survey (N = 27)



Abbreviation: Q, question.

Participants considered the software’s “allowance for minimum disclosure” and “comprehensive privacy protection that is not currently available” as the most important MiNDFIRL benefits in the final online survey (Figure 18). “Required checks to ensure privacy protection” and the “potential of misuse by authorized users” (eg, negligence, not sufficient training) were among participants greatest concerns, with mean scores of 3.5 and 3.4, respectively (Figure 18). The Likert responses for all 11 additional information options were “necessary” (4 choices), “very necessary” (6 choices), or “essential” (1 choice) (Figure 18).

These qualitative data and the subsequent participant rankings helped us identify the key issues and concerns that patients care about in secondary database research. This information was used to develop the initial draft FAQ document that was used in study C.2. The prioritized MiNDFIRL issues, concerns, and benefits identified in the study C.1 NGT informed the content for each question and answer in the FAQ as well as how that information was presented in the initial draft.

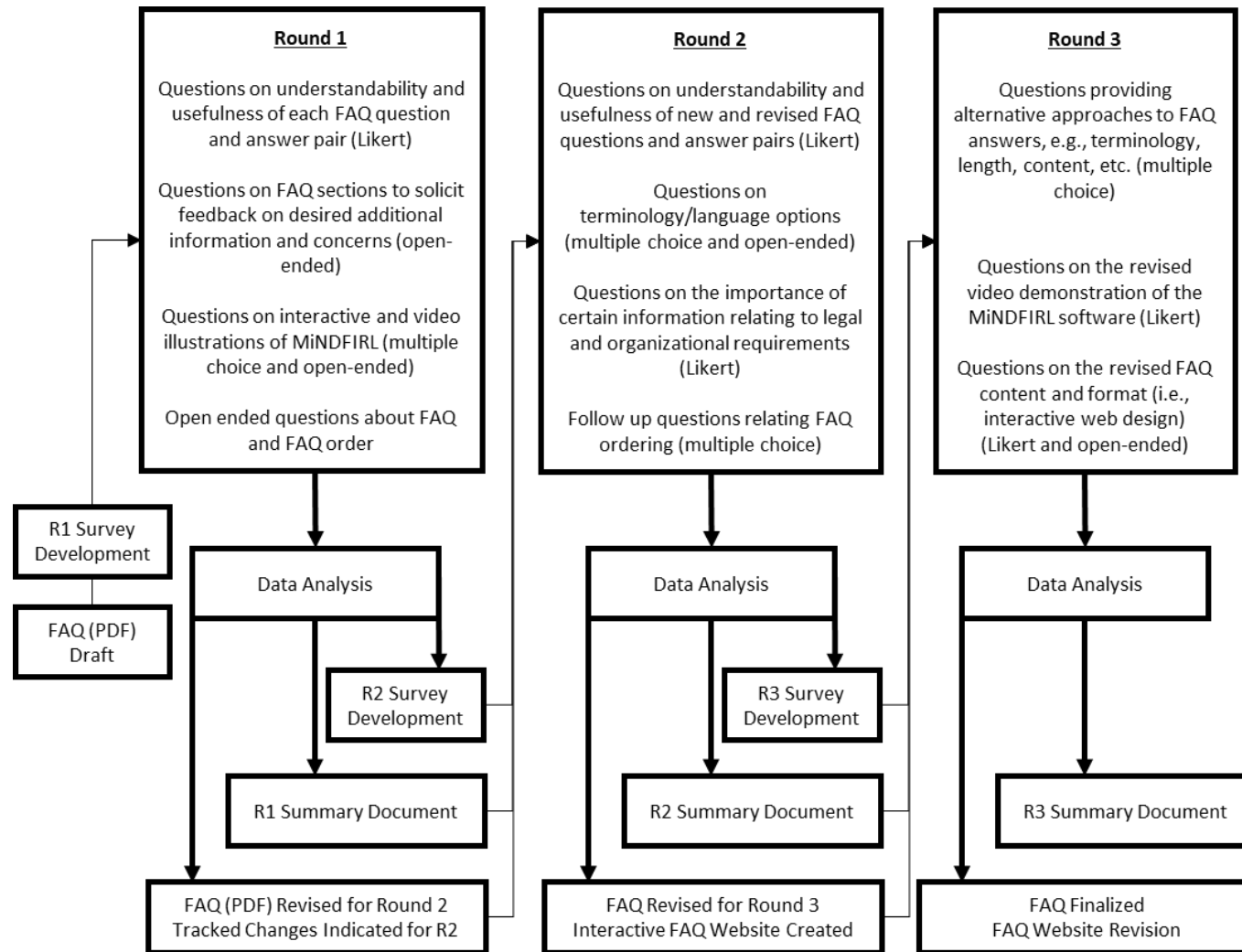
Study C.2 – patient Delphi process.

Objective. Study C.2 used the findings of study C.1 (as well as studies D.1 and D.2) to create a draft FAQ document that would be refined through a Delphi process to enhance communication and trust between patients and researchers.

Methods

Design. We conducted a 3-round Delphi study using a web-based questionnaire (ie, Qualtrics). Figure 19 depicts the overview of the process. Delphi facilitates anonymous and confidential feedback from patients with diverse perspectives without the biases common to other consensus techniques such as group discussions and interviews.^{43,44} The Delphi approach is particularly suited to investigate communication strategies with patients and data subjects, where there are differences in thought and the body of knowledge is still developing.^{43,45}

Figure 19. Overview of Delphi Process and Round Content



Abbreviations: FAQ, frequently asked questions; MiNDFIRL, MInimum Necessary Disclosure For Interactive Record Linkage; R, round.

Reproduced from Schmit et al. *J Med Internet Res.* 2020;22(12):e20783. <https://doi.org/10.2196/20783>. Reprinted with permission from the Journal of Medical Internet Research (Copyright ©2020). All Rights Reserved.

We used patients with chronic conditions for the Delphi panel because they are likely to have conditions of interest to secondary database researchers, and their data are likely to be dispersed among multiple health care providers (ie, requiring RL).

We asked participants questions about a draft FAQ in each Delphi round, revising the FAQ between rounds based on participants' feedback. The FAQ drafts included visual images and a short video demonstrating the MiNDFIRL RL software.

Participants. We recruited patients using purposive sampling from PCORnet and from employees and staff of Texas A&M University via email listservs. We included participants with at least 1 diagnosed chronic disease and with at least 2 physician visits for their condition in the previous year. We compensated participants \$100 with gift cards on a graduated basis. Only those who completed the prior Delphi round were invited to participate in the next round.

Delphi round procedure. Each Delphi round contained a mix of open-ended questions and 5-point Likert scale questions, asking for feedback on the FAQ, including whether FAQ sections provided information that was understandable or important to patients. All 3 surveys were pilot tested by members of the research team before administration. Survey rounds 2 and 3 focused on FAQ language and content where participant feedback suggested divergent opinions. In rounds 2 and 3, we also provided participants with a summary of the participant feedback from the previous round and a "redline" version of the revised FAQ. Participants were given just over a week to complete each round, with 2 reminder emails.

After each Delphi round, we triaged FAQ sections as consensus or nonconsensus. FAQ sections with negative feedback from fewer than 3 participants (ie, "disagree" or "strongly disagree") were deemed consensus language. We made only minor edits to consensus sections (eg, terminology revisions). FAQ sections that received negative feedback from ≥ 3 participants were deemed nonconsensus; we made substantial revisions to these sections and solicited additional feedback in subsequent rounds. We note that our consensus criteria required positive or neutral feedback from between 90% and 92%, depending on the round, which is

highly conservative and notably higher than that with other Delphi studies (eg, simple majorities).^{43,46-48}

Rounds 2 and 3 contained additional questions to identify preferred terminology and strategies for communicating key concepts to the patient community. If participant feedback suggested a divergence of opinions, we explored the divergence and solicited specific feedback in subsequent rounds (eg, providing alternative approaches to answering an FAQ question).

After all 3 rounds, 2 researchers conducted a preliminary inductive thematic analysis of the open-ended responses to provide additional context for our primary result: the FAQ template document.

Results. The principal result of this study is the final FAQ template document. This final FAQ document was built on the foundation provided by the study C.1 NGT and refined through this Delphi study. The final FAQ template had a Flesch-Kincaid readability score of 8.66, slightly higher than the initial FAQ score of 7.82.

Thirty-eight (86%) participants completed round 1 (Table 17). Females (72%) and non-Hispanic White individuals (87%) were disproportionately represented. Participants were between 21 and 78 years old, with a mean age of 50 years. A total of 37/38 (97%) participants completed round 2, and 33/37 (86%) participants completed round 3, giving a total 73% (33/45) completion rate for invited participants.

Round 1. In round 1, participants responded favorably to the overall FAQ, with 90% of responses being within the strongly agree and agree categories across all 50 item questions on the FAQ, while only 10% of responses were either neutral or negative. Large majorities of participants strongly agreed or agreed that the FAQ questions were easy to understand (93%) and that the answers provided in the FAQ were easy to understand and contained useful information (87%).

The question with the strongest negative feedback (18 participants either disagreed or strongly disagreed that the FAQ item contains useful information) was, “What will you do if you

Table 17. Demographic Information of Participants Who Completed Round 1 (N = 38)

Characteristics	Values
Age (years)	
Median (SD)	49 (14.6)
Range	21-78
Gender, n (%)	
Male	10 (26)
Female	28 (74)
Education, n (%)	
Some college credit/no degree	3(8)
Associate degree	7(18)
Bachelor's degree	11(29)
Masters' degree	11(29)
Doctoral degree	6 (16)
Race & Ethnicity, n (%)	
White	33(87)
African American/Black	3(8)
Asian	1(3)
Other	1(3)
Avg physician visit in 12 months, n (%)	
2-5 times	15(39)
6-10 times	14(37)
>10 times	9(24)
Self-reported health status, n (%)	
Excellent	2(5)
Very good	10(26)
Good	14(37)
Fair	12(32)

Reproduced from Schmit et al. *J Med Internet Res.* 2020;22(12):e20783. <https://doi.org/10.2196/20783>. Reprinted with permission from the Journal of Medical Internet Research (Copyright ©2020). All Rights Reserved.

discover that my data has been misused?” Many participants objected to the lack of specificity with the provided answer.

Only 7 FAQ question/answer items were designated nonconsensus and triaged for substantial edits for round 2. In round 1, five participants (13%) had concerns about the terminology used to describe information subsets (ie, “identifiers” and “nonidentifiers”).

Round 2. Round 2 had fewer questions (25 items) than did round 1 due to a high level of agreement and positive feedback. Generally, revisions related to terminology and readability. Several new FAQ sections and visual aids were added to improve understanding of specific concepts (eg, “What is Patient Matching?”).

Of the newly created and revised nonconsensus FAQ items, 90% of participants agreed that the questions and answers were easy to understand, and 89% agreed that the FAQ items contained useful information. Negative feedback of the length and complexity often directly conflicted with positive feedback relating to detail and clarity.

Round 3. In round 3, we provided alternative options for 2 nonconsensus FAQ items. The majorities of participants (57% and 60%) favored the shorter alternatives. However, when provided an opportunity for substantial cuts (ie, simplification), a strong majority (25 of 33) of the participants preferred including the information for those who wanted it.

We attempted to address conflicting participant feedback relating to the competing values of simplicity and brevity vs detail and completeness by changing the FAQ format in a few ways. First, we created an interactive FAQ website with expandable sections. Second, we replaced definitions in the main text with definition pop-up boxes that appear when a user’s mouse hovers over key terms. Third, we bolded important text within each FAQ section to aid content skimming. A strong majority of the round 3 participants (31 of 33) found this revised format helpful or very helpful.

Preliminary thematic analysis. The preliminary inductive thematic analysis identified 9 themes in the participants’ open-ended responses (Table 18).

Table 18. Themes of Participant Feedback

Theme	Brief description of feedback
Simplicity and brevity	Preference for short and direct explanations
Detail and completeness	Preference for complete explanations with sufficient details for clarity and understanding
Readability	Preference for content that is easy to read and uses layman language
Terminology and definitions	Preference for clearly defined terms and avoiding technical jargon
Tone	Feedback concerning the tone of explanations, eg, conversational, not patronizing
Examples	Feedback concerning the utility of examples of key concepts
Visuals	Feedback concerning the utility of graphics, video, and interactive aids
Data disposition and future uses	Feedback concerning patient concerns relating to what happens to the research data at the end of the project, eg, destruction, reuse, storage
Patient rights	Feedback concerning the explanation of patient rights and protections

Study C.3 – online survey with a sample representative of the US population.

Objective. Study C.3 evaluated the template FAQ language developed in study C.2 using a large sample representative of the US population.

Methods. We conducted an online survey in February 2020 to obtain feedback from the general US population. The survey was administered through Qualtrics, and potential participants were identified through a private company (Dynata) which conducts online surveys and national sampling. Participants were compensated consistent with the company’s policies. We included participants who were adult US residents fluent in English.

We provided participants with background information about RL, our software, and the purpose of this study. Participants were then given access to the online FAQ and were asked to answer 3 questions related to the document. These were the following:

Do you prefer this FAQ format to a traditional privacy statement that you might read on a website, mobile app, or computer program? (Likert scale)

How useful is the FAQ document? (Likert scale)

Please provide and comments, opinions, and suggestions about the FAQ document you reviewed in the section below. (open ended)

Results. The online survey was sent to 687 people. Of those, 470 respondents completed the survey (response rate = 68.4%) without detected quality issues (eg, rapid click-through). Generally, we met our census sampling targets for gender, race/ethnicity, age, income, and census region (Table 19) for a sample closely representative of the US population. However, the sample does differ from Census targets on education attainment (eg, 32.6% vs 51.0% high school or less, respectively).

Most participants found the FAQ document useful, with 82% indicating that it was extremely, very, or moderately useful. Just over half of respondents (51%) preferred the FAQ to a traditional privacy statement (39% had no preference). Statistical analysis of the results by sociodemographics found no statistical differences between race, gender, education, income, or region. However, participants who had a chronic condition found the FAQ document more useful ($P < .001$) and had a stronger preference for the FAQ format ($P < .05$) than did those who did not have a chronic condition. We present selected results by race, education, and chronic condition in Figure 20.

The FAQ document received both positive and negative feedback in the open-ended questions. Positive open-ended feedback included comments on the ease of navigating the FAQs, the use of a patient-centered voice, and the detailed and comprehensive explanations. Negative open-ended feedback included comments suggesting there was too much information, concerns over privacy risk, and comments requesting more detailed information (ie, study and institution-specific information, like breach protocols). Participants suggested 2 improvements: adding a search function and providing the document in more languages. Examples are quotes are given in Figure 21. Figure 22 is an example screenshot of the FAQ.

Table 19. Sociodemographic, Clinical Characteristics, and Privacy Attitude Scores of Participants

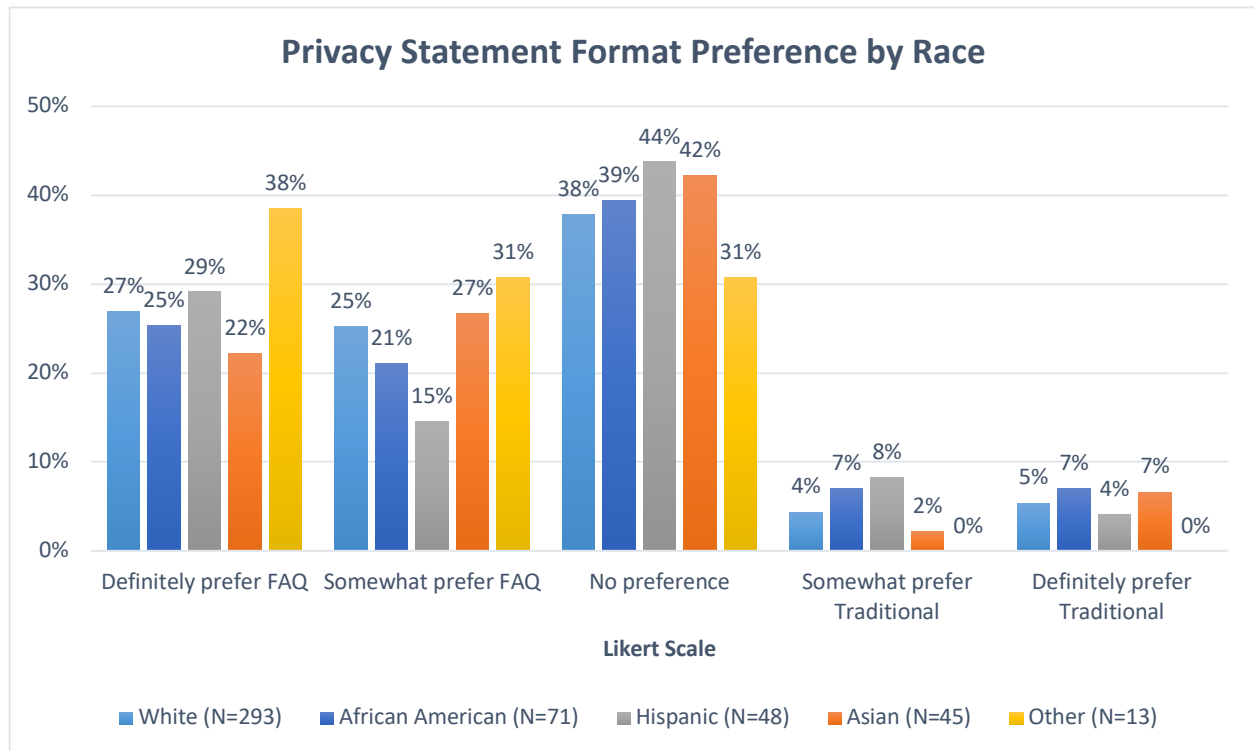
Participant characteristic	N = 470	Sample target ^{a,b}
Age category, %		
18-24 y	9.4	13.1
25-34 y	17.4	17.5
35-44 y	20.9	17.5
45-54 y	21.1	19.2
55-64 y	14	15.6
≥65 y	17.7	17.2
Gender, %		
Male	43.8	48.5
Female	55.7	50.5
Other/prefer not to answer	0.4	-
Race category, %		
White	62.3	63.7
African American	15.1	12.2
Hispanic	10.2	16.4
Asian	9.6	4.7
Other	2.8	3.0
Income category, %		
\$30 000 or less	31.9	
\$30 000-\$59 999	26.2	
\$60 000-\$99 999	18.7	
≥\$100 000	23.2	
Educational level, %		
High school or less	32.6	51.0
Some college or college degree	58.9	31.0
Master's and PhD/doctoral	8.5	
Region, %		
Midwest	19.4	22.0
Northeast	23.6	18.2
South	34.7	36.2
West	22.3	23.6

Health insurance coverage, % ^b		
Private	34	64.7
Medicare	23	17.7
Medicaid	16.8	17.9
Uninsured	9.8	8.5
VA/TRICARE	2.1	3.6
Multiple	14.3	14.5
Any chronic condition, %		
No	63.8	
Yes	36.2	

^aSurvey sampling targets based on Census data.

^bInsurance data were not used as a sampling target. These data show 2018 insurance statistics from the US Census for survey sampling comparisons.⁴⁹ Our survey solicited mutually exclusive responses, in contrast to the US Census data, which do not exclude persons with multiple insurance types from these groups.

Figure 20. Privacy Statement Format and Usefulness Preference



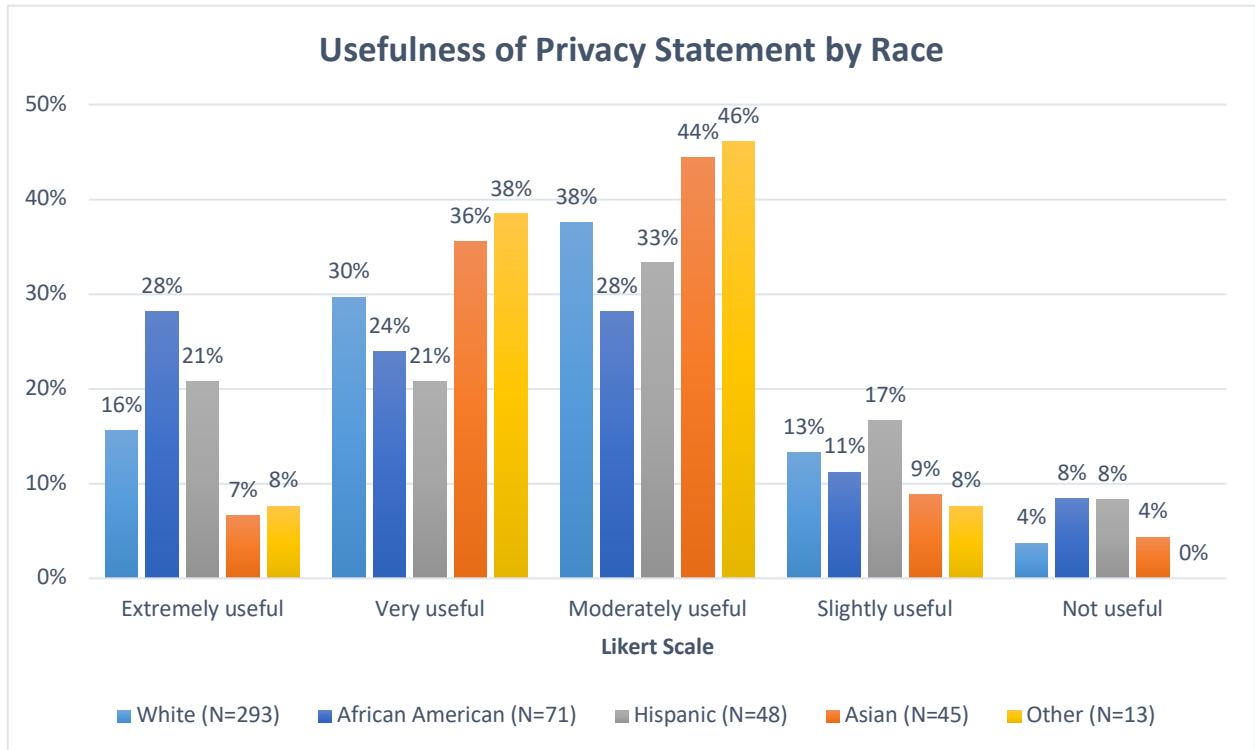
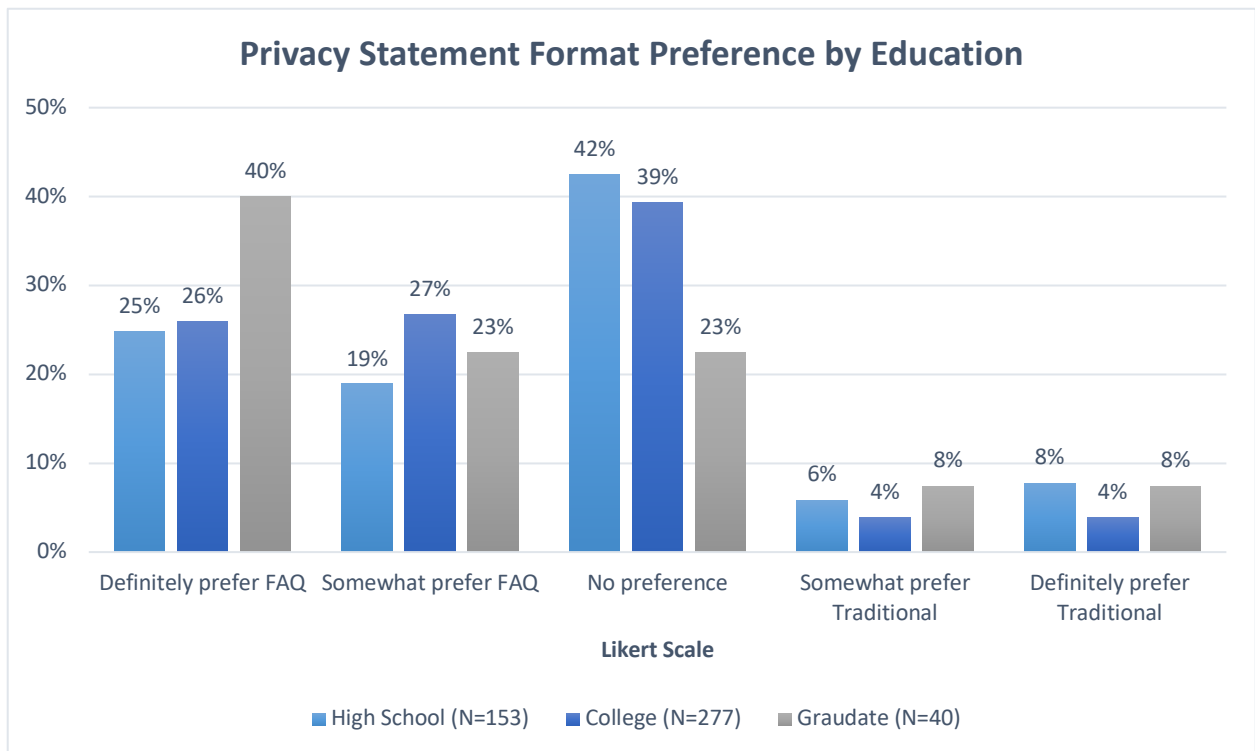


Figure 20. Privacy Statement Format and Usefulness Preference (cont'd)



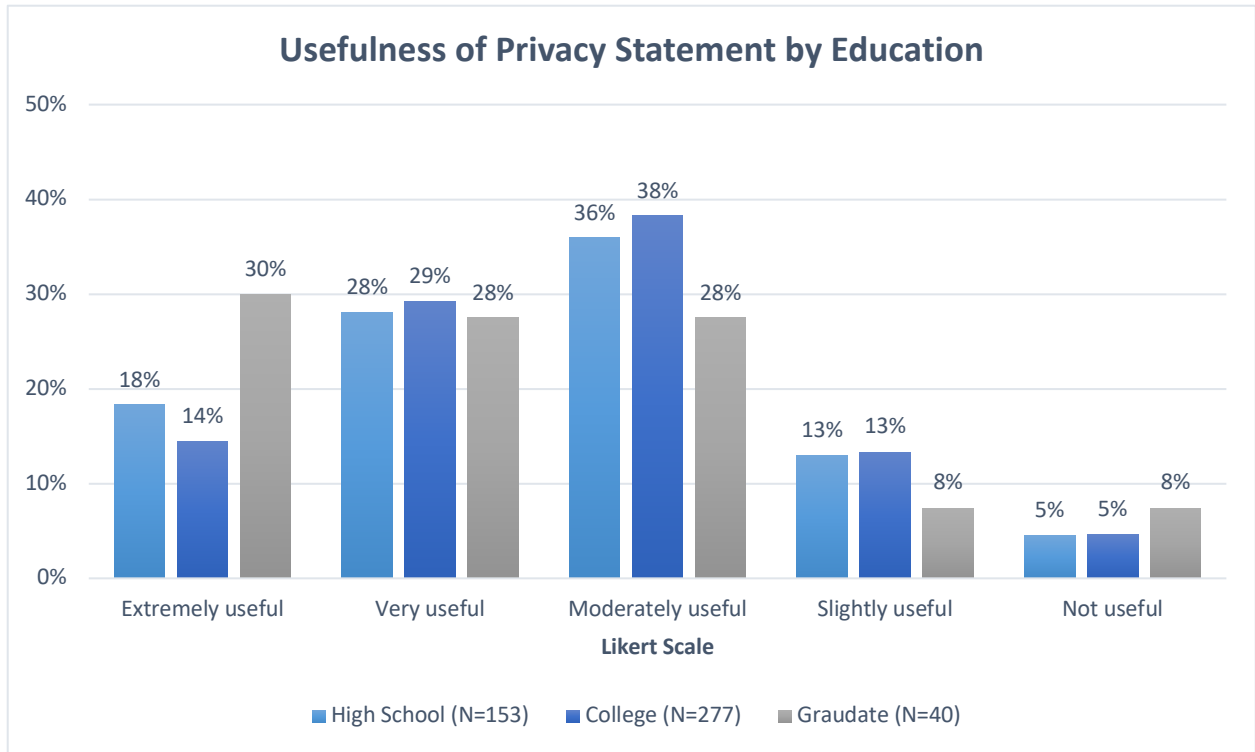
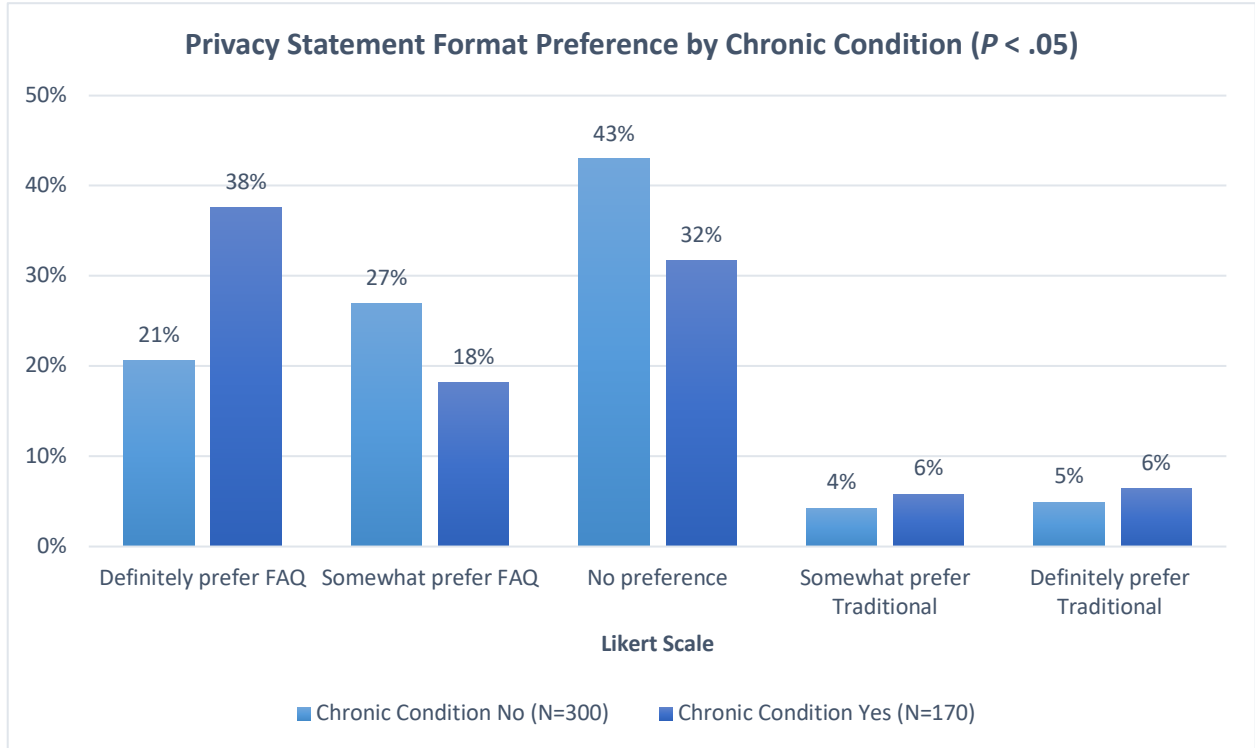
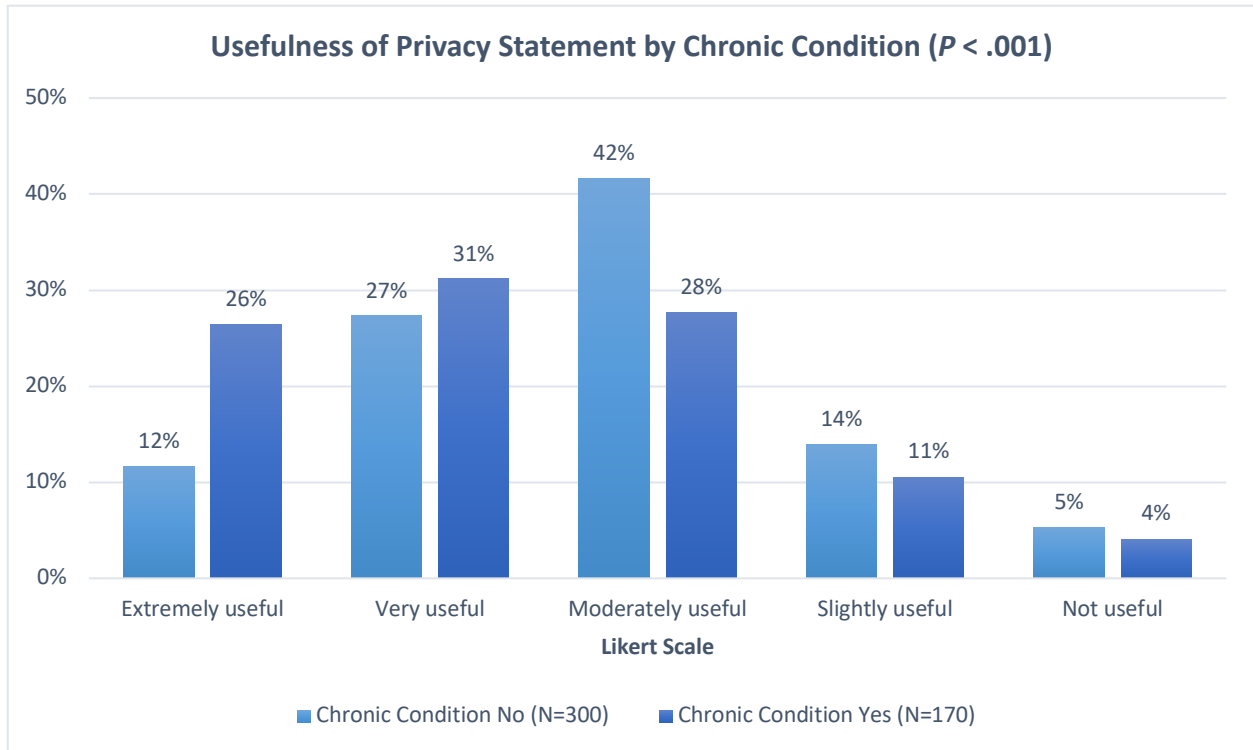


Figure 20. Privacy Statement Format and Usefulness Preference (cont'd)





Abbreviation: FAQ, frequently asked question.

Figure 21. Example Quotes for Positive and Negative Feedback

Example quotes for positive feedback

“I really like the FAQ layout because it’s not as cumbersome to read as a traditional privacy policy. It’s easier to open up each section as I like.”

“Definitely like the sections being broken apart into questions I might have. I think it reframes the document into a user-centered POV and I think that shows consideration.”

“I like how thorough this FAQ is and the in-depth responses. I also like being able to choose the topics that most interest me, or that I have less an understanding of.”

“I like the way it is set up, it is easy to follow and navigate. It might be nice if there was a search box since there is so much information, it could take a while to find the exact answer you are looking for.”

“Make sure it is available in multiple languages. Otherwise it looks fine to me.”

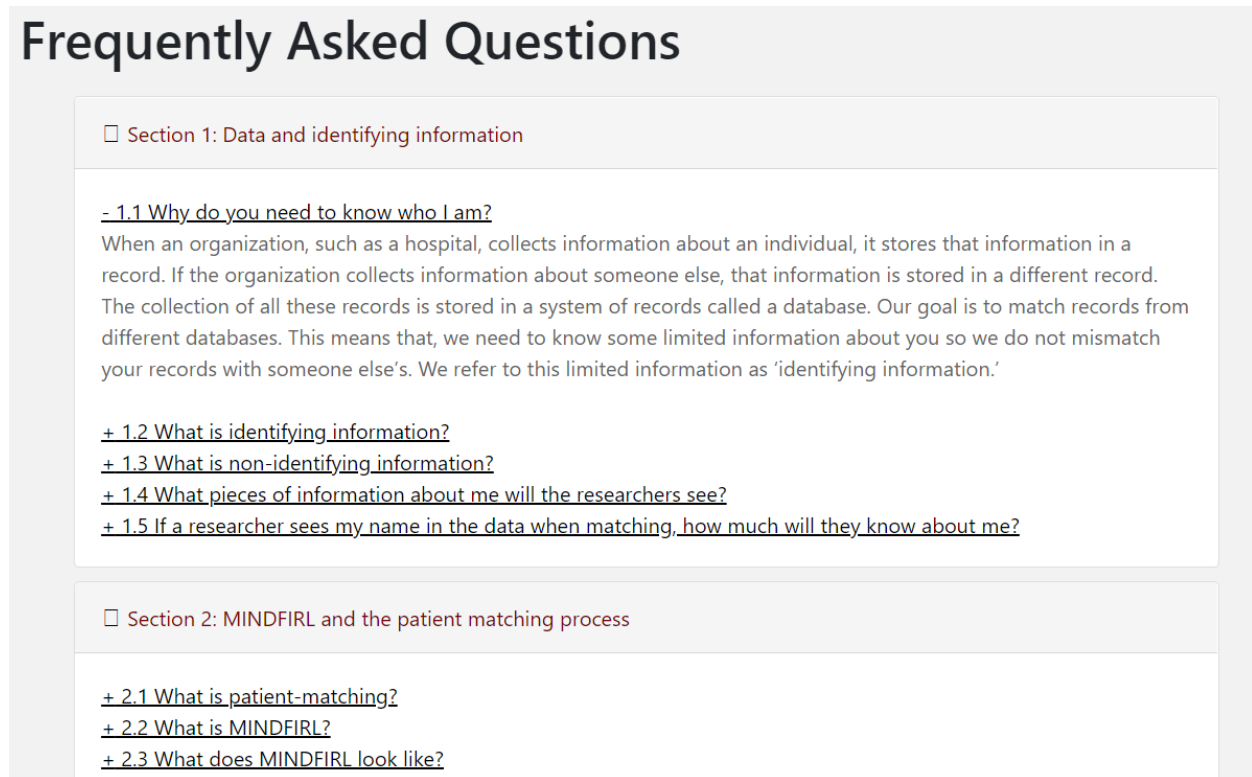
Example quotes for negative feedback

“I think they are just too long and no one will actually read them.”

“While I like the format, it doesn’t change the problem of a system getting compromised. So it’s helpful in providing answers, but would not eliminate my concerns. Best intentions don’t always lead to good results.”

“I would like to see what would happen if there would be a data breach. How would a company be accountable.”

Figure 22. Example FAQ Screenshot



Aim 3.2 – IRB Application Template for MiNDFiRL

The primary objective of aim 3.2 was to create a template IRB application and response language to use when communicating with IRBs about conducting secondary database research using the MiNDFiRL software. To accomplish this, we conducted 2 studies (D.1 and D.2) utilizing NGT and Delphi methods with participants who were ELSI experts. Below, we describe the objective, methods, and results from the 2 studies.

Study D.1 – ELSI NGT.

Objective. Study D.1 was designed to acquire the perspectives of ELSI experts, including IRB professionals, on the perceived benefits and risks of MiNDFiRL and to identify what information is important for IRB research determinations (eg, exemption). The findings from this research informed the development of the IRB application template and template responses to assist researchers using MiNDFiRL to communicate effectively about the software

with an IRB. Portions of the content below have been presented at the AcademyHealth 2018 Annual Research Meeting.⁵⁰

Methods. We conducted 2 NGT sessions (in person and online). We recruited ELSI experts and IRB professional attendees at the 2017 Advancing Ethical Research Conference for the in-person NGT session with assistance from the conference organizers. We compensated participants with a \$25 gift card and gave 1 participant an additional \$100 gift card at random. We facilitated the online session in January 2018, recruiting through professional networks like Public Responsibility in Medicine and Research. Online participants were given a \$20 gift card and entered a raffle for an additional \$50.

With AC input, we drafted 3 questions for the NGT sessions that would solicit useful information for creating a template IRB application and responses. The questions were:

- “What do you perceive as the benefits of using the MiNDFIRL approach for database record linkage?”
- “What do you perceive as the risks for subjects of data when using the MiNDFIRL approach for database record linkage?”
- “For research using the MiNDFIRL approach for record linkage, what other information would you need to know if you were serving on the IRB as the public representative for reviewing and approving an IRB application?”

These questions are closely related to the issues germane to the ethical review of research conducted by IRBs.

Each session contained a 15-minute online tutorial of how the MiNDFIRL software would operationalize RL across various databases and gave participants hands-on experience using MiNDFIRL to link records across 2 databases.

Each NGT session followed a 3-step structure which was completed in 45 to 60 minutes. In the first phase, participants were given a total of 30 minutes (10 minutes per question) to individually build a list of responses to each question. In the second phase, the research team

gathered and organized the list of the responses. The facilitator led group discussions for each question, sharing participant responses and seeking clarification as appropriate. Common responses were combined by the participants into broader themes. At the end, participants were asked to vote on the 2 most important themes per question in ranked order (primary and secondary).

After both studies were completed, 4 researchers identified the highest-ranked themes across both groups. We emailed all participants the combined list and asked them to identify their top 2 themes by importance (primary and secondary) to build consensus across groups.

Results. Eleven participants completed both phases of our study: the initial NGT session (5 participated in person, 6 participated online) and the combined final voting. Participants' professional affiliations included IRB, compliance, and ELSI research positions.

Both groups generated 34 total responses for all 3 questions. Importantly, there was general consensus about most issues raised, and similar and overlapping themes suggested saturation. After removing or combining responses with matching thematic content, 13 total responses remained (5, 4, and 4 responses for questions 1, 2, and 3, respectively) (Table 20).

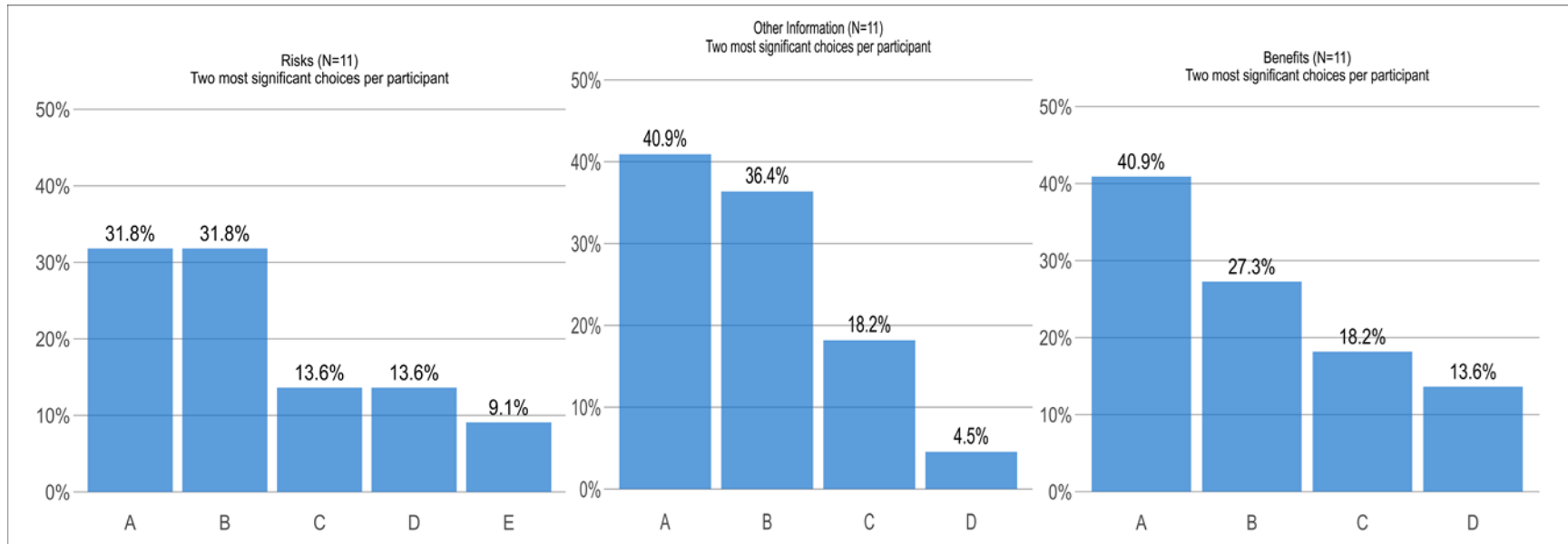
Figure 23 depicts participant priorities. The highest ranked benefits were the potential to facilitate research and the potential to promote responsible data use and good governance. The highest ranked concerns were the potential to enable flawed research (eg, inaccuracy from user or software errors) and inadequate organizational administrative controls. The highest ranked additional information for an IRB review was evidence of MiNDFIRL RL validity and information on administrative controls. Overall, we received constructive feedback from all participants about the importance of our research.

Table 20. Emerging Themes From ELSI Experts

Question 1:
What do you perceive as the benefits of using the MINDFIRL framework for database record linkage?
<i>Options:</i>
A. Potential to facilitate the execution of “research protocols” (eg, providing a tool for researchers to link data, deidentify data, and reidentify data)
B. Potential that the framework will promote “responsible and accountable data use and good data governance”
C. Potential to facilitate research and data sharing “approval processes”
D. Potential to reduce the “risk of disclosure”
Question 2:
What do you perceive as the risks for subjects of data when using the MINDFIRL framework for database record linkage?
<i>Options:</i>
A. Potential that software will enable flawed research (eg, linking flawed data or enabling research that uses inaccurately linked data from user or software errors)
B. Possibility that an organization’s administrative controls (ie, training and user rules) permit inappropriate use of the software
C. Potential for unnecessary privacy loss/identity exposure to authorized personnel (ie, among researchers)
D. Potential for privacy loss to unauthorized personnel (eg, hacking)
E. Potential for a lack of accountability for disclosures
Question 3:
For research using the MINDFIRL framework for record linkage, what other information would you need to know if you were serving on the IRB as the public representative for reviewing and approving an IRB application?
<i>Options:</i>
A. Evidence for the validity of record linkage when using the software
B. The administrative controls (eg, organizational rules, policies, and required training) and data governance structure that would be used under MINDFIRL
C. The nature of software security and vulnerability issues, if any
D. Specific details regarding the nature of the data used for record linkage

Abbreviations: ELSI, ethical, legal, and social implications; MiNDFIRL, MiNimum Necessary Disclosure For Interactive Record Linkage.

Figure 23. Consensus Votes of Emerging Themes From ELSI Experts



Abbreviation: ELSI, ethical, legal, and social implications.

Study D.2 – ELSI Delphi process.

Objective. Study D.2 builds on the study D.1 findings by using the Delphi technique with ELSI experts to create and refine a template IRB application for secondary database research and template responses that researchers can use when communicating with IRBs.

Methods. We conducted a 3-round online Delphi study. We recruited a panel of ELSI experts, particularly IRB professionals by consulting university, hospital, and Veterans Affairs websites and at the 2018 Advancing Ethical Research Conference and 2018 American Public Health Association Annual Meeting (which overlapped in dates and venue). We included participants at least 18 years of age with English fluency and a professional ELSI role. We provided graduated payments amounting to \$100 for all 3 rounds. A total of 18 ELSI experts participated in our study.

Three research team members (each with legal, ethical, or IRB experience) drafted the initial IRB template, incorporating feedback from study D.1 and our advisory board. We then designed and pilot tested our Delphi instrument with 4 ELSI experts on our board.

In each round, participants were asked to provide feedback on the IRB template, including whether the sections provided information needed to approve future studies using the MiNDFIRL software. All Delphi rounds contained a mix of open-ended questions and 5-point Likert scale questions related to the IRB application template and template responses. The research team reviewed and discussed all proposed revisions in response to participant feedback after the completion of each Delphi round. We gave participants just over a week to complete each Delphi round. We sent 2 reminder emails to reduce attrition.

In response to round 1 feedback, we gave participants draft MiNDFIRL training documents and access to a hands-on MiNDFIRL tutorial³⁰ in round 2. In rounds 2 and 3, we provided participants with a summary of the prior round's feedback and a redline version of the revised IRB template to enable them to easily identify changes and associated feedback. Rounds 2 and 3 contained questions exploring diverging or conflicting feedback. Although the Delphi technique does not demand a specific consensus threshold, our consensus criteria cutoff

was negative feedback by 3 or fewer individuals, which was conservative and higher than that from previous Delphi studies.^{44,51}

Results. Of the 18 ELSI experts, 17 fully completed round 1 (94.4%), 15 completed round 2 (88.2%), and 13 completed round 3 (86.7%), for an overall response rate of 72.2% (13/18). The mean (SD) age of all participants was 41 (10.2) years, 88.2% were females, and 47.1% and 29.4% had master's or bachelor's level education, respectively. The majority were certified IRB professionals (82.4%); all were IRB staff, and 70.6% were also IRB members. In round 1, no participants reported having a fundamental concern described in the IRB template such that they would never approve the described research. Most respondents (76.5%) were completely or partially satisfied with the privacy protection provisions, and 70.6% deemed the template language on administrative controls as essential to approve future research. About 59% of the respondents were extremely confident or confident that the description of the risk precautions and monitoring plans were sufficient and appropriate, and 41.2% were moderately confident. Strong majorities of respondents wanted additional information on the storage of linked data for future use (82.4%) and additional information about who will receive the data and what data will be shared and how (100%). In total, 82.4% of participants indicated that the prototype IRB application provided sufficient information for an IRB determination of a database-only study involving RL. Major suggestions included revising the language of some IRB questions and proposed PI responses.

In round 2, fourteen of the 15 participants (93.3%) found the MiNDFIRL training and tutorial to be very useful or useful, and 1 found it to be somewhat useful. At least 12 out of the 15 round 2 participants indicated that the revisions and language changes were improvements. Minor revisions were suggested for the protocol description and privacy protections.

In round 3, six of the 13 participants (46.2%) were extremely satisfied with the revised IRB application template, and 53.8% were somewhat satisfied. Nine of the 13 ELSI experts strongly agreed or agreed with the statement that the use of the MiNDFIRL software will further reduce risk to the minimum necessary to conduct reliable RL; no respondents disagreed.

Aim 3.3 – DUA Template for MiNDFIRL

Methods. Legal agreements for sharing and using data are critically important to ensure that data are used and protected appropriately and data rights are respected. The contents of a legal agreement can vary considerably based on party priorities, negotiations, and other important context. Importantly, laws often require legal agreements to contain specific terms or address specific issues. Consequently, the negotiating parties must often adapt legal terms and conditions to the law(s) applicable to their data.^{52,53} For this reason, there cannot be a universal legal agreement template for sharing or using data in all contexts. Nevertheless, template legal language can be a useful starting place for negotiations between data-sharing partners and to reduce transactional friction between parties.

For this reason, we developed template language for a DUA to accompany the MiNDFIRL software. Our objective was to identify and address key issues to parties using MiNDFIRL during the RL process in template language that would also be useful for parties to adapt for their own purposes. Accordingly, this DUA template was influenced by the findings from the aim 3 studies (ie, studies C.1, C.2, C.3, D.1, and D.2), which helped identify issues of concern and points of confusion that could be addressed in the template language or highlighted as issues for consideration.

Given that MiNDFIRL is an RL software, it was important for us to consider that different data sets used in the RL process might be governed by different data protection laws and effectuated by different organizational policies. To make this template language as broadly relevant as possible, we chose to adapt an agreement that relates to data maintained by the US Department of Health and Human Services (DHHS) and protected by the Privacy Act of 1974.⁵⁴ The Privacy Act, which regulates systems of records maintained by the federal government, is one of the most broadly applicable data protection laws in the US legal data protection framework (although federal agencies have some flexibility in implementing the Privacy Act requirements). Consequently, this DUA template is likely to be somewhat easier to adapt to data projects linking data from different sectors (eg, linking data from the Department of

Housing and Urban Development with DHHS data) than is a DUA that was designed to address a more narrow data protection framework (eg, the HIPAA of 1996).

In creating this template language, we paid close attention to existing language⁵⁵ issued by DHHS for DUAs and solicited comments from attorneys and database researchers. While we expect that parties and their legal counsels will adapt this template language to their individual and organizational needs, it is important that the template language adheres to provided best-practice guidance.

Results. The principal result of this study is the final DUA template document that has been released with the MiNDFIRL software on the GitHub repository as well as the project website.⁵⁶ Full references are also available in the Appendix.

DISCUSSION

Minimum Necessary Standard and Practical Challenges

Research on information privacy has shown the complex balance of providing protection while still allowing utility from the legitimate use of personal data for social benefit.^{32,57} Among the core principles for designing privacy-enhanced systems is to limit disclosures of protected information to only those necessary for achieving a given purpose. This principle is central to various data protection laws in the form of *minimum necessary* or *need-to-know* information disclosure standards. Laws like HIPAA, the Privacy Act of 1974, the confidentiality protections for substance abuse disorder records in 42 CFR Part 2, as well as many state laws (eg, California Consumer Privacy Act of 2018) use similar legal standards to permit legitimate uses of data while protecting privacy by limiting extraneous disclosure.⁵⁸⁻⁶¹ Similarly, the EU General Data Protection Regulation uses the principle of “data minimisation” to limit data use to what is necessary for a permitted purpose.⁶²

Moreover, minimizing data disclosures is good practice for ethical data use and stewardship. Beyond the legal requirements and organizational policies that formally restrict data uses, trust and relationships between potential partners are essential to removing barriers to data sharing.^{53, 63} Many organizations will—often rightly—refuse to release data to parties without an existing healthy relationship with the parties despite the presence of permissive laws or organizational policies. Tools that can support minimizing data disclosures while promoting transparency and accountability in data sharing can help foster trust between parties by providing assurances of responsible data use.³⁹ These assurances are critical to address lingering confidentiality concerns that often impede permitted data uses.

However, practically implementing a process for sharing protected data that restricts disclosures to the minimum necessary is a daunting task.^{18,22,53,64} It is rare that a data project knows exactly which data elements and observations are needed ahead of time. Instead, data science is often an iterative process of learning from the data and refining the analysis until useful results are obtained. Moreover, the iterative nature of analytic methods also means that

the required data dynamically change over the course of the project. Practically, in many situations, it is the case that all the data are decided to be the “minimum necessary.”⁶⁵

These dynamics can lead to serious consequences when negotiations and legal agreements must be made (eg, DUAs) between different organizations for data sharing. Perceptions about what constitutes the minimum necessary can differ between data-sharing partners, leading to prolonged project delays.⁶⁶ Even worse, funded projects may be cancelled when researchers are not able to pass a vetting process for giving full access to protected data.⁶⁶ One reason for this is because there are no practical tools to facilitate data disclosures that closely meet the minimum necessary legal standards, opening negotiations to potentially lengthy debates about what information is truly “necessary” or even potentially necessary.

Synergies With Related Research in RL

The need to obtain full access to large amounts of sensitive data is especially true in RL CER projects. Linking data sets to address broader questions in CER usually requires obtaining approvals necessary to gain full access to all PII. This often poses unacceptable levels of disclosure both legally and ethically.⁶³ Thus, there are ongoing efforts to use encryption-based, privacy-preserving RL algorithms when there is a common reliable identifier (eg, SSN or insurance ID), with most of the efforts on building technology to securely share the encryption key (eg, hash key).⁶⁷ However, the validity of the results is fully reliant on the quality of the common identifier. The biggest challenge in this line of research is that the validity of the common identifier in any specific project cannot be verified. There are also continued efforts to improve the hashing (ie, encryption) algorithms to better handle approximate matching in RL.⁶⁸ Others have experimented with privacy-enhanced methods that use indirect identifiers for RL.⁶⁹

Notably, all of these privacy-protecting RL approaches work on the computational problem called “private RL.” Private RL focuses on linking data securely given a predetermined linkage function (eg, same SSN), which is unknown in most real applications.⁷⁰⁻⁷² Human interaction is required to determine the linkage function, clean and standardize the data, and tune the automatic models.³³ Yet, most theoretical private RL approaches do not support these required human interactions with the data. Our research complements efforts in private RL by

developing a privacy-enhanced interactive data integration framework for researchers to directly, but securely, integrate person-level data to support valid, reliable, and replicable research in CER.

For example, 1 major issue in using any data linked using private RL methods is that there is limited understanding of the validity of the linked data, making it very difficult to interpret the findings in CER studies. This issue is illustrated well in a recent pilot study that attempted to link 4 PPRN registries with 14 health plans to confirm patient-reported clinical conditions from the insurance database.⁷³ The researchers were able to link 21% of the PPRN members with various confirmation rates by clinical condition ranging from 75% for multiple sclerosis to 50% for rheumatoid or psoriatic arthritis. Rates improved to 93% and 67%, respectively, for members with more than 5 years of continuous health plan enrollment, but the results were based on less than 5% linkage rate. The overlap between the 21 616 PPRN members and the 14 health plans is unknown, making it difficult to assess if the matching rate is acceptable and how to interpret the clinical conditions of 79% (no restriction on period of enrollment) or 95% (≥ 5 years of enrollment) of the nonmatched members. Such uncertainties in real data are a fundamental property of data science, making it even more critical to pursue research methods that can bound these uncertainties so that the results can support decisions and actions. Our research is complementary to these private-RL efforts and can be used on the data to estimate a linkage rate and provide ways to bound the uncertainty with only minimal disclosure of PII.

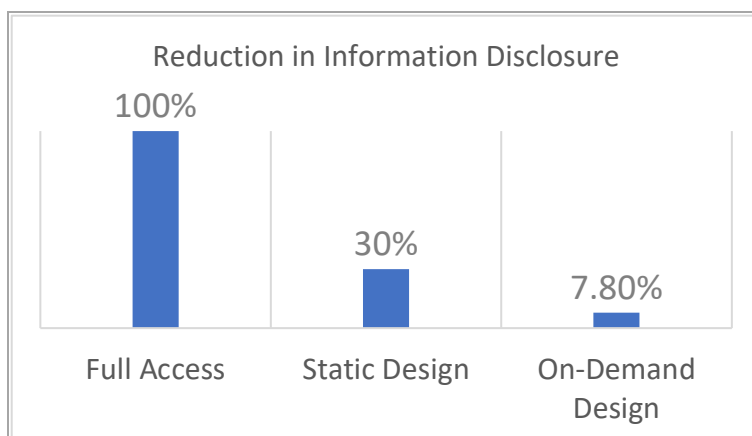
Our Contributions to Privacy-Enhancing Technology Development for CER

We address the absence of practical tools to facilitate the minimum necessary data disclosures by investigating the software design elements that can facilitate privacy-enhanced secondary data analysis. Our research contributes a novel interactive interface where we start with fully masked deidentified data and let users click to open when more information is required for good decisions. The interface is meant to serve as a complement to algorithmic methods for detecting possible duplicates or discrepancies among similar records. For uncertain cases requiring human review and judgment, the system presents the flagged pairs as rows in a

tabular interface with different data fields separated by columns (see Figure 4). The system takes advantage of 3 techniques to enhance privacy protection: (1) minimum necessary disclosure via just-in-time, incremental information access; (2) transparent accountability by quantifying the privacy risk due to the disclosure made; and (3) limiting access to data during linkage via a budget.

Our findings clearly demonstrate that privacy-by-design is effective in implementing the minimum necessary standard in RL. The studies for aims 1 and 2 investigated these design elements by evaluating the trade-offs between the privacy and utility of identifying information for human decision-making. Our results demonstrate that with well-designed software, it is possible to greatly limit the amount of identifying information available to human decision makers without negatively affecting utility or human effectiveness in RL. The static design in study A.1 with visual masking was able to reduce information disclosure to only 30%, and the dynamic just-in-time design in study A.2 was able to further reduce this to only 7.8% with no real impact on RL decision-making (Figure 24).

Figure 24. Reduction in Information Disclosure While Maintaining Human Decision-Making in RL



Abbreviation: RL, record linkage.

Most importantly, our work provides strong evidence that with sound research and attention to software design, there is a way to find the “sweet spot” where with minimum information disclosure we can obtain high-quality linked data for CER. Feedback from the

expert reviews supports the notion that an intermediate level of access other than “all or nothing” can provide better accuracy than can no access but more protection than full access. The main feedback from the experts was that the system facilitated safe linkage, providing more confidentiality to patients without compromising on the quality of the results, which provides a good balance between the “all or none” access to PII. Some experts had concerns about the potential increase in time required for using the system. However, although there were slight increases in completion time for some interventions of our study, no statistically significant differences in completion times were found among the different modes, perhaps because preventing users from looking at details that are not needed to increase privacy streamlines the interface so that they are not inundated with too much information. This is likely to reduce the time needed to complete the data task. Thus, the selective disclosure not only has the benefit of significantly reducing privacy risk, it may also enhance data workers’ attention to their task.

However, the findings in both formative studies also show there is a limit to how much data can be hidden before negatively influencing the quality of judgment in RL decisions. This supports the common understanding that information privacy is a budget-constrained problem. Previous research has demonstrated that effective information privacy requires reasoning about the trade-off between privacy and utility for a given context.⁷⁴⁻⁷⁷ Consequently, there is no “one-size-fits-all” solution, and there is no way to benefit from using data without taking some privacy risks. Essentially, researchers must think through the benefit of obtaining high-quality linked data for CER and decide on acceptable levels of data disclosure to achieve this goal while using the best-designed software to maximum privacy protection through the privacy-by-design approach.

The main challenge in these privacy-utility trade-off decisions in leveraging data for societal benefit is the complexity of the decision about how much privacy to trade off for better data quality and who has the authority to make these decisions. These decisions have direct impact on rapidly leveraging data in public health emergencies (eg, the COVID-19 pandemic) for social benefit and should be well thought out. Yet, most often these decisions are a by-product

of the labyrinth of legal restrictions, many organizational policies, and individual decisions by data custodians, leading to suboptimal case-by-case decisions balancing societal benefit and individual interests (eg, privacy⁵²).

Thus, in aim 3, our broad goal was to engage with diverse stakeholder groups (1) to educate on issues of RL and information privacy (2) to make informed consensus decisions on the social benefit of linking data, key software designs that would maximize the privacy protection and public trust, and (3) to transparently communicate on any remaining potential risks to personal privacy when using data without consent in CER. The aim 3 findings supported the cocreation of 3 documents: the template FAQ privacy statement that transparently discloses critical information to the public to promote transparency and trust in research; the template IRB application and responses that help researchers communicate these critical issues to ethical review bodies to speed review processes; and the template DUA that helps research institutions negotiate with data custodians and owners, reducing transaction frictions to obtain new data. Together, we anticipate that these documents will significantly reduce the transactional and startup challenges present in new secondary database research projects.

Our findings from numerous focus groups and surveys with patients provide insights into which research-related information is useful to patients and how researchers can communicate such information in patient voice. The results indicate several patient communication considerations, including that (1) patients have diverse and varied preferences; (2) tone is important but challenging; and (3) patients want information on security, identifiers, and final disposition of information. These findings align with the current understanding of health literacy and its challenges.^{78,79} Communication is essential to transparency and ethical data use, yet it is exceedingly challenging. Developing FAQ template language to accompany complex software may enable researchers to provide greater transparency when informed consent is not possible.

Similarly, our findings from engaging with the ELSI experts provide insight into issues of human subjects research in secondary database CER. Most experts (10 of 13) agreed that the use of the MiNDFIRL software will further reduce risk to the minimum necessary when

conducting reliable RL. Experts unanimously reported that IRB applications that describe research protocols using the specialized software need to report whether the information technology (IT) department reviewed and approved MiNDFIRL. The majority deemed information related to who maintains, reviews, and has access to the software (12 of 13) and which IT department is managing the server with the software and PII (11 of 13) as important for inclusion in the IRB applications. More broadly, all participants supported the notion of privacy-utility trade-off that despite the use of the MiNDFIRL software, all database studies have an inherent risk of unexpected disclosures due to a potential breach of the computer system hosting the data. Nonetheless, most experts (11 of 13) noted that subjects in the database-only studies experience no greater risks than the risks experienced in ordinary life. The concept of risks in ordinary life is important because it is the main criterion for assessing risk in human subjects research. Challenges arise when new risks to ordinary life are introduced with new technologies that become widespread (eg, smartwatches). Our expert panel supported the notion that being included in a database has become an ordinary risk, as today, most of our daily activities are already digitized. Thus, most importantly, our ELSI panel supported (1) the determination of large-database-only studies as minimal risk studies under the Common Rule and (2) the use and effective communication of well-designed software to minimize the privacy risk in large-database studies.

Study Limitations

While the formative studies contributed empirical evidence of the potential benefits of the developed techniques along with knowledge of potential trade-offs between decision-making and privacy, we note that there are limitations to making broad generalized statements when the participants were not fully randomized, as in the case of our formative studies (although the 5 study groups in study A.1 were balanced along 4 variables). In addition, although the combination of expert reviews and summative evaluation provides valuable feedback from more knowledgeable and complex contexts, it is challenging to conduct larger systematic evaluations with expert populations. Thus, the summative evaluations with practitioners have thus far included 12 data workers and 2 case studies at 2 institutions. This limitation is common in evaluations of software systems, and the methods used in this research

and report follow the general guidelines of human-computer interaction (despite this limitation, the peer-review paper describing these findings was nominated for a best paper award at the top forum in the field).¹⁷ The availability of experts and professional data workers is limited. Therefore, we are limited in our ability to claim the extent to which the results of the formative studies apply to other (and different) real-world settings. Though our summative evaluation did consider multiple sites and different data linkage challenges with different data sources, the chosen subset of 2 sites is still limited. Findings and feedback from the sample of data workers are likely to vary among individuals, as are the personal opinions of different experts and specifics of data work. Additional studies across other sites and data sources would strengthen the knowledge of generalizability and transfer the potential for different operational data environments. The summative evaluation also revealed novel insights about the importance of individual user differences in personality, work style, and background that can influence the use of the designed software features. Further research is still needed to understand the necessary training for effectively preparing different types of data workers to benefit from the developed techniques. Another limitation is attrition in both Delphi studies. Only 73% of participants invited to participate in study C.2, and only 72% of participants invited to participate in study D.2 completed all 3 rounds of those studies. We cannot determine how those participants would have evaluated the final documents (ie, template FAQ and template IRB documents).

All of our studies in aim 3, with the exception of the large-scale survey of the FAQ (study C.3), have the inherent limitation of qualitative research, which is not to generalize the findings beyond the study population. In addition, our NGT focus groups and Delphi surveys were selected as methods to purposefully and iteratively build consensus among the study participants. We achieved high consensus in all of our studies, but the results do not include the full range of divergent opinions of all groups. Thus, our results should be interpreted in the context of consensus building among the study participants, with their associated characteristics as experts, as provided in the Methods and Results section. Importantly, the ELSI expert panel group had diverse professional representation (eg, IRB program director, IRB member, and compliance and ELSI research positions) with appointments in academic,

governmental, hospital, or health system settings. More limited in scope were the patient participants of studies C.1 and C.2, who were recruited heavily from PPRNs, many of whose members are highly engaged in research. The study population included diverse patients in terms of gender, race (Black, Hispanic, and Asian), and education level (high school graduate or equivalent, some college, up to PhD). However, more than 50% were White, female, and well educated (college or graduate school), raising some concerns of representativeness. Nonetheless, in study C.3, where we ran a large-scale survey with a nationally representative sample, there were no statistical differences in preference for the FAQ format or the usefulness of the FAQ document in terms of race, gender, or education, providing stronger support for the cocreated FAQ document.

Future Research

There are several directions for future research. First, the project scope did not include investigating automated algorithms required for a comprehensive hybrid human-computer system. We had prior experience building machine learning–based RL models that we used throughout the study, but more research is needed to (1) make these codes usable by non–machine learning experts, (2) incorporate the automatic RL component into a full end-to-end system, (3) build RL benchmarking systems that can compare the different model results, and (4) investigate how to incorporate privacy-preserving RL–based methods using hashed data. In terms of software design, further research is needed to conduct a formative study to investigate the trade-off between using easy-to-understand functions (eg, percentage of information disclosed) and more accurate but complex functions (eg, KAPR score) for quantifying the privacy risk. More human-computer interaction research is needed to assess the understandability and interpretation of risk given the dependence on a person’s data literacy and understanding of uncertainty. Research can similarly iterate on knowledge of how to represent uncertainty and risk through different methods of textual, numerical, and visual explanations of numerical values for the context of the RL parameters. The most immediate enhancement to MiNDFIRL in future research should be to expand the pair display to implement group display such that multiple records that are potentially very similar can be grouped and viewed together. Investigating an effective interface design for manual review of a

group of records will require user studies similar to those done in this project for designing the pair interface. Finally, further analysis of the evaluations that were conducted, such as (1) quantifying potential biases in the PII that were clicked/viewed and (2) quantifying the heterogeneity in the EHR data and patient-generated data, may reveal more useful insights.

In aim 3, our work was one of the first to engage widely with stakeholders to educate and build consensus decisions on complex issues of privacy and the use of person-level data (without individuals' consent) for CER. More work is needed in database research to develop specialized software and to communicate the relevant details of the software and infrastructure to gain public trust. Investigations into how to expand the community consultation model to leverage newer online technology may allow for obtaining broader engagement than the current status quo (ie, informed consent waiver). Such procedures may augment the current IRB process, allowing for broader support of database studies for public benefit.

CONCLUSIONS

Prior research has demonstrated the detrimental effect of not allowing sufficient human participation in data tasks such as RL.^{12, 14,15,80-83} Errors that are not properly managed in machine-only data linkage propagate to subsequent data analyses. Thus, to obtain high-quality data and bias-free RL for research, human involvement is essential to fine-tune the results from automated systems (eg, parameter settings, setting cutoff thresholds, iterative data standardization, building training data sets, validating results).^{33,83} Therefore, to produce accurate linkages, some identifying data, under some suitable conditions, must be revealed to trusted persons.

Our research provides evidence that incremental, partial disclosure of identifying data can be highly effective for ensuring compliance with the principle of revealing the “minimum necessary” data needed to link records from the same person and transparent access to data. Our controlled experiments demonstrate that properly designed software support can reduce the amount of identifying information needed to make a correct match. The results also suggest limits to how much data can remain undisclosed and still support high-quality decisions. Findings from our ELSI expert panel further supported this notion of privacy-utility trade-off that, despite the use of the MiNDFIRL software, all database studies will have inherent risks (eg, system breaches). ELSI experts agreed that large-database-only studies were minimal-risk studies under the Common Rule, and the use and effective communication of well-designed software such as MiNDFIRL were important to minimize the privacy risk in such studies. Moreover, in our experience, cocreating public privacy statements with patients in the patient voice can support transparency and improve patient trust.

Based on these findings, we iteratively designed, implemented, and released the open-source MiNDFIRL prototype software along with 3 companion documents describing the use of MiNDFIRL for high-quality RL to support CER/PCOR. To our knowledge, this is the first open-source software to include template documentation to facilitate transparent communication. The research and findings presented in this report demonstrate the potential for privacy enhancement in research that requires linking individual patient data from ≥ 2 large

observational data sets as well as opportunities for future research to further improve the approach.

REFERENCES

1. Clark RE, Samnaliev M, Baxter JD, Leung GY. The evidence doesn't justify steps by state Medicaid programs to restrict opioid addiction treatment with buprenorphine. *Health Aff (Millwood)*. 2011;30(8):1425-1433.
2. Fisher WH, Clark R, Baxter J, Barton B, O'Connell E, Awew G. Co-occurring risk factors for arrest among persons with opioid abuse and dependence: Implications for developing interventions to limit criminal justice involvement. *J Subst Abuse Treat*. 2014;47(3):197-201.
3. Boscoe FP, Schrag D, Chen K, Roohan PJ, Schymura MJ. Building capacity to assess cancer care in the Medicaid population in New York State. *Health Serv Res*. 2011;46(3):805-820.
4. Bradley CJ, Given CW, Luo Z, Roberts C, Copeland G, Virnig BA. Medicaid, Medicare, and the Michigan Tumor Registry: a linkage strategy. *Med Decis Making*. 2007;27(4):352-363.
5. DuVall SL, Fraser AM, Rowe K, Thomas A, Mineau GP. Evaluation of record linkage between a large healthcare provider and the Utah Population Database. *J Am Med Inform Assoc*. 2012;19(e1):e54-e59.
6. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Med Care*. 2002;40(8 Suppl):IV3-IV18.
7. Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: a survey. *IEEE Trans Knowl Data Eng*. 2006;19(1):1-16.
8. Kang H, Getoor L, Shneiderman B, Bilgic M, Licamele L. Interactive entity resolution in relational data: a visual analytic tool and its evaluation. *IEEE Trans Vis Comput Graph*. 2008;14(5):999-1014.
9. Köpcke H, Thor A, Rahm E. Evaluation of entity resolution approaches on real-world match problems. *Proceedings VLDB Endowment*. 2010;3(1-2):484-493.
10. Newcombe HB, Kennedy JM, Axford S, James AP. Automatic linkage of vital records. *Science*. 1959;130(3381):954-959.
11. Joffe E, Byrne MJ, Reeder P, et al. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. *J Am Med Inform Assoc*. 2014;21(1):97-104.

12. Bronstein JM, Lomatsch CT, Fletcher D, et al. Issues and biases in matching Medicaid pregnancy episodes to vital records data: the Arkansas experience. *Maternal Child Health J.* 2009;13(2):250-259.
13. Ilangovan G, Ramezani M, Kum H-C. A benchmarking system to evaluate the effectiveness and efficiency of machine learning algorithms for record linkage. *MS Thesis Texas A&M University Department of Computer Science and Engineering.* Presented at: AMIA 2020 Virtual Annual Symposium; November 16, 2020; Virtual. <https://oaktrust.library.tamu.edu/handle/1969.1/186390>
14. Baldi I, Ponti A, Zanetti R, Ciccone G, Merletti F, Gregori D. The impact of record-linkage bias in the Cox model. *J Eval Clin Pract.* 2010;16(1):92-96.
15. Dusetzina SB, Tyree S, Meyer A-M, Meyer A, Green L, Carpenter WR. *Linking Data for Health Services Research: a Framework and Instructional Guide.* Report No. 14-EHC033-EF. Agency for Healthcare Research and Quality. September 2014. Accessed January 14, 2022. <https://www.ncbi.nlm.nih.gov/books/NBK253313/>
16. Lahiri P, Larsen MD. Regression analysis with linked data. *J Am Stat Assoc.* 2005;100(469):222-230.
17. Ragan ED, Kum H-C, Ilangovan G, Wang H. Balancing privacy and information disclosure in interactive record linkage with visual masking. Paper presented at: 2018 CHI Conference on Human Factors in Computing Systems; April 21-26, 2018; Montreal, Québec, Canada. Accessed January 13, 2022. <https://dl.acm.org/doi/10.1145/3173574.3173900>
18. Kum H-C, Ragan ED, Ilangovan G, Ramezani M, Li Q, Schmit C. Enhancing privacy through an interactive on-demand incremental information disclosure interface: applying privacy-by-design to record linkage. Paper presented at: Fifteenth USENIX Conference on Usable Privacy and Security; August 11-13, 2019; Santa Clara, CA. Accessed January 13, 2022. <https://dl.acm.org/doi/10.5555/3361476.3361489>
19. Goth G. Running on EMPI. Health information exchanges and the ONC keep trying to find the secret sauce of patient matching. *Health Data Manag.* 2014;22(2):52, 54, 56 passim.
20. McCoy AB, Wright A, Kahn MG, Shapiro JS, Bernstam EV, Sittig DF. Matching identifiers in electronic health records: implications for duplicate records and patient safety. *BMJ Qual Saf.* 2013;22(3):219-224.
21. US Government Accountability Office. *Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information.* April 1, 2001. <https://www.gao.gov/assets/gao-01-126sp.pdf>

22. Gostin LO, Levit LA, Nass SJ. *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. National Academies Press; 2009.
23. Morris G, Farnum G, Afzal S, Robinson C, Greene J, Coughlin C. *Patient Identification and Matching Final Report*. Office of the National Coordinator for Health Information Technology. Published February 7, 2014.
https://www.healthit.gov/sites/default/files/patient_identification_matching_final_report.pdf
24. Hickam D, Totten A, Berg A, Rader K, Goodman S, Newhouse R. *The PCORI Methodology Report*. Patient-Centered Outcomes Research Institute (PCORI). Published November 2013. <https://www.pcori.org/assets/2013/11/PCORI-Board-Meeting-Methodology-Report-for-Acceptance-1118131.pdf>
25. Population Informatics Lab. hybridRL. May 20, 2020. Accessed July 22, 2021.
https://github.com/pinformatix/hybridRL_code_and_models
26. Population Informatics Lab. rLErrorGenerator. May 20, 2020. Accessed July 22, 2021.
<https://github.com/pinformatix/rLErrorGenerator>
27. Population Informatics Lab. Privacy preserving interactive record linkage (PPIRL) via information suppression. June 7, 2020. Accessed September 30, 2020.
<https://pinformatix.org/ppirl/>
28. Population Informatics Lab. MINDFIRL. April 7, 2020. Accessed September 30, 2020.
<https://github.com/pinformatix/mindfirl>
29. Fleurence R, Selby JV, 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(2):393-400.
30. Population Informatics Lab. Study of record linkage and information disclosure. June 7, 2020. Accessed September 30, 2020. <http://mindfil4.herokuapp.com/introduction>
31. Population Informatics Lab. Frequently asked questions. June 7, 2020. Accessed September 30, 2020. <https://pinformatix.org/ppirl/faq/faq.htm>
32. Narayanan A, Shmatikov V. Myths and fallacies of “personally identifiable information.” *Comm ACM*. 2010;53(6):24-26.
33. Kum H-C, Krishnamurthy A, Machanavajjhala A, Reiter MK, Ahalt S. Privacy preserving interactive record linkage (PPIRL). *J Am Med Inform Assoc*. 2014;21(2):212-220.
34. Shneiderman B, Plaisant C. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson Education India; 2010.

35. Li Q, D'Souza AG, Schmit C, Kum H-C. Increasing transparent and accountable use of data by quantifying the actual privacy risk in interactive record linkage. *arXiv*. Preprint posted online June 7, 2019. <https://arxiv.org/abs/1906.03345>
36. Acquisti A, Adjerid I, Balebako R, et al. Nudges for privacy and security: understanding and assisting users' choices online. *ACM Comput Surv*. 2017;50(3):1-41.
37. Ur B, Kelley PG, Komanduri S, et al. How does your password measure up? The effect of strength meters on password creation. Paper presented at: 21st USENIX Conference on Security Symposium; August 8-10, 2012; Bellevue, WA. Accessed January 14, 2022. <https://dl.acm.org/doi/10.5555/2362793.2362798>
38. Giannouchos TV, Ferdinand AO, Ilangovan G, et al. 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. *J Am Med Inform Assoc*. 2021;28(8):1746-1755.
39. Schmit C, Ajayi K, Ferdinand A, 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(12):e20783. doi:10.2196/20783
40. Horton J. Nominal group technique: A method of decision-making by committee. *Anaesthesia*. 1980;35(8):811-814.
41. Harvey N, Holmes CA. Nominal group technique: an effective method for obtaining group consensus. *Int J Nurs Pract*. 2012;18(2):188-194.
42. Bouchard TJ Jr, Hare M. Size, performance, and potential in brainstorming groups. *J Appl Psychol*. 1970;54(1 Pt 1):51-55.
43. Williams PL, Webb C. The Delphi technique: a methodological discussion. *J Adv Nurs*. 1994;19(1):180-186.
44. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs*. 2000;32(4):1008-1015.
45. Adler M, Ziglio E. *Gazing Into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health*. Jessica Kingsley Publishers; 1996.
46. Butterworth T, Bishop V. Identifying the characteristics of optimum practice: findings from a survey of practice experts in nursing, midwifery and health visiting. *J Adv Nurs*. 1995;22(1):24-32.
47. Hasson F, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs*. 2000;32(4):1008-1015. doi:10.1046/j.1365-2648.2000.t01-1-01567.x

48. Powell C. Myths and realities of the Delphi technique. *J Adv Nurs*. 2003;41(4):376-382.
49. Berchick ER, Hood E, Barnett JC. *Health Insurance Coverage in the United States: 2017. Current Population Reports*. US Census Bureau. Published September 2018.
<https://www.census.gov/content/dam/Census/library/publications/2018/demo/p60-264.pdf>
50. Giannouchos T, Kum H-C, Ferdinand AO, Schmit C, Ilangovan G, Ragan ED. Patients' and stakeholders' perceptions of risk and benefits of the privacy preserving interactive record linkage (PPIRL) framework. *Academy Health*; 2014.
https://pinformatics.org/ppirl/img/ppirl_poster.pdf
51. Butterworth T, Bishop V. Identifying the characteristics of optimum practice: findings from a survey of practice experts in nursing, midwifery and health visiting. *J Adv Nurs*. 1995;22(1):24-32.
52. Hulkower R, Penn M, Schmit C. Privacy and confidentiality of public health information. In: Magnuson J, Dixon B, eds. *Public Health Informatics and Information Systems*. Springer; 2020:147-166.
53. Schmit C, Kelly K, Bernstein J. Cross sector data sharing: necessity, challenge, and hope. *J Law Med Ethics* 2019;47(2_suppl):83-86.
54. Centers for Medicare & Medicaid Services. Instructions for completing the disproportionate share hospital data use agreement (DUA). Department of Health and Human Services; 2020. Accessed September 30, 2020.
<https://www.cms.gov/Medicare/CMS-Forms/CMS-Forms/Downloads/CMS-R-0235D2.pdf>
55. Department of Health and Human Services. Department of Health and Human Services enterprise performance life cycle framework: practices guide, data use agreement. 2020. Accessed September 30, 2020.
https://www.hhs.gov/sites/default/files/ocio/eplc/EPLC%20Archive%20Documents/55-Data%20Use%20Agreement%20%28DUA%29/eplc_dua_practices_guide.pdf
56. Population Informatics Lab. MiNDFIRL (Minimum Necessary Disclosure For Interactive Record Linkage). June 7, 2020. Accessed September 30, 2020.
<https://pinformatics.org/ppirl/mindfirl.php>
57. Wartenberg D, Thompson WD. Privacy versus public health: the impact of current confidentiality rules. *Am J Public Health*. 2010;100(3):407-412.
58. *In re Estate of Broderick*. 34 Kan. App. 2d 695, 703, 125 P.3d 564, 570 (2005).

59. Guthrie J. Time is running out-the burdens and challenges of HIPAA compliance: a look at preemption analysis, the minimum necessary standard, and the notice of privacy practices. *Ann Health Law*. 2003;12(1):143-177, V.
60. *v S. US Department of Veterans Affairs*, 218 F.R.D. 619, 631 (E.D. Wis. 2003), amended on reconsideration in part, 222 F.R.D. 592 (E.D. Wis. 2004).
61. West's Ann Cal Civ Code § 1798.100, et seq (2021).
62. Team IP. *EU General Data Protection Regulation (GDPR)*. IT Governance Limited; 2017.
63. Schmit C, Giannouchos T, Ramezani M, Zheng Q, Morrissey MA, Kum HC. US privacy laws go against public preferences and impede public health and research: survey study. *J Med Internet Res*. 2021;23(7):e25266. doi:10.2196/25266
64. Kum H-C, Krishnamurthy A, Machanavajjhala A, Ahalt SC. Social genome: putting big data to work for population informatics. *Computer*. 2013;47(1):56-63.
65. Department of Health and Human Services. Health Information Privacy. Does the HIPAA Privacy Rule strictly prohibit use, disclosure, or request of an entire medical? If not, are case-by-case justifications required each time the entire medical record is disclosed? December 19, 2002. Updated July 26, 2013. Accessed September 30, 2020. <https://www.hhs.gov/hipaa/for-professionals/faq/213/what-conditions-may-health-care-provider-use-entire-medical-record/index.html>
66. van Panhuis WG, Paul P, Emerson C, et al. A systematic review of barriers to data sharing in public health. *BMC Public Health*. 2014;14(1):1144.
67. Kho AN, Cashy JP, Jackson KL, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. *J Am Med Inform Assoc*. 2015;22(5):1072-1080.
68. Kuzu M, Kantarcioglu M, Durham EA, Toth C, Malin B. A practical approach to achieve private medical record linkage in light of public resources. *J Am Med Inform Assoc*. 2013;20(2):285-292.
69. Patient-Centered Outcomes Research Institute. *How Do Patients Feel About Sharing and Linking Health Data for Research?* 2020. Accessed September 30, 2020. <https://www.pcori.org/research-results/2014/how-do-patients-feel-about-sharing-and-linking-health-data-research>
70. Hall R, Fienberg SE. Privacy-preserving record linkage. In: Domingo-Ferrer J, Magkos E, eds. *Privacy in Statistical Databases. PSD 2010. Lecture Notes in Computer Science, vol 6344*. Springer; 2010:269-283.

71. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. *BMC Med Inform Decis Mak*. 2009;9(1):41.
72. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. *Inf Syst*. 2013;38(6):946-969.
73. Agiro A, Chen X, Eshete B, et al. Data linkages between patient-powered research networks and health plans: a foundation for collaborative research. *J Am Med Inform Assoc*. 2019;26(7):594-602.
74. Kum H-C, Ahalt S. Privacy-by-design: understanding data access models for secondary data. *AMIA Jt Summits Transl Sci Proc*. 2013;2013:126-130.
75. Li T, Li N. On the tradeoff between privacy and utility in data publishing. Paper presented at: 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; June 28-July 1, 2009; Paris, France. Accessed January 14, 2022. <https://dl.acm.org/doi/10.1145/1557019.1557079>
76. Kifer D, Machanavajjhala A. No free lunch in data privacy. Paper presented at: 2011 ACM SIGMOD International Conference on Management of Data; June 12-16, 2011; Athens, Greece. Accessed January 14, 2022. <https://dl.acm.org/doi/10.1145/1989323.1989345>
77. Dinur I, Nissim K. Revealing information while preserving privacy. Paper presented at: Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems; June 9-11, 2003; San Diego, CA. Accessed January 14, 2022. <https://dl.acm.org/doi/10.1145/773153.773173>
78. Kindig DA, Panzer AM, Nielsen-Bohlman L. *Health Literacy: a Prescription to End Confusion*. National Academies Press; 2004.
79. Health Resources and Services Administration. Health literacy. US Department of Health Human Services. Reviewed August 2019. Accessed May 25, 2020. <https://www.hrsa.gov/about/organization/bureaus/ohe/health-literacy/index.html>
80. Lane J. *Optimizing the Use of Micro-Data: an Overview of the Issues*. SSRN; August 2005. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=807624
81. Lane J, Schur C. Balancing access to health data and privacy: a review of the issues and approaches for the future. *Health Serv Res*. 2010;45(5 Pt 2):1456-1467.
82. Lahiri P, Larsen MD. Regression analysis with linked data. *J Am Stat Assoc*. 2005;100(469):222-230.

83. Kum H-C, Ahalt S, Pathak D. Privacy-preserving data integration using decoupled data. In: Altshuler Y, Elovici Y, Cremers A, et al, eds. *Security and Privacy in Social Networks*. Springer; 2013:225-253.

RELATED PUBLICATIONS

Published (Submitted in Appendix)

- Kum H-C, Ragan E, Ilangovan G, Ramezani M, Li Q, Schmit C. Enhancing privacy through an interactive on-demand incremental information disclosure interface: applying privacy-by-design to record linkage. Paper presented at: SOUPS'19: Fifteenth USENIX Conference on Usable Privacy and Security; August 11-13, 2019; Santa Clara, CA. Accessed January 13, 2022. <https://dl.acm.org/doi/10.5555/3361476.3361489>.
- Ragan ED, Kum H-C, Ilangovan G, Wang H. Balancing privacy and information disclosure in interactive record linkage with visual masking. Paper presented at: 2018 CHI Conference on Human Factors in Computing Systems; April 21-26, 2018; Montreal, Québec, Canada. Accessed January 13, 2022. <https://dl.acm.org/doi/10.1145/3173574.3173900>. *CHI2018 Honourable Mention Best Paper Award (top 5% of all submissions)*. Also presented at the 14th Symposium on Usable Privacy and Security (SOUPS) Aug 2018 as invited poster.
- Li Q, D'Souza AG, Schmit C, Kum H-C. Increasing transparent and accountable use of data by quantifying the actual privacy risk in interactive record linkage. *arXiv*. Preprint posted online June 7, 2019. <https://arxiv.org/abs/1906.03345>
- Schmit C, Ajayi KV, 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(12):e20783. doi:10.2196/20783
- Giannouchos T, Ferdinand AO, Ilangovan G, et al. 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. *J Am Med Inform Assoc*. 2021;28(8):1746-1755.

Draft Completed to Be Submitted in the Near Future

- Ferdinand AO, Giannouchos T, Schmit C, Kum H-C. Engaging with IRB experts about benefits and risks of using privacy-preserving software: a case study of Minimum Necessary Disclosure for Interaction Record Linkage (MiNDFIRL). Planned submission to *J Med Ethics*.

ACKNOWLEDGMENTS

We thank the following staff and collaborators for their important contribution to this work: Kobi Ajayi, Elmer Bernstam, Jeffery Curtis, Adam D'Souza, Theodoros Giannouchos, Gurudev Ilangovan, Qinbo Li, Benjamin Nowell, Mahin Ramezani, and Han Wang.

APPENDIX: OUTCOME

MiNDFIRL (Minimum Necessary Disclosure For Interactive Record Linkage)

Project Website:

<https://pinformatics.org/ppirl/index.php>

Open Source Software:

- (1) On project website: <https://pinformatics.org/ppirl/mindfirl.php>
- (2) On GitHub: <https://github.com/pinformatics/mindfirl>

Three Companion Documents

- (1) A template privacy statement: This is a dynamic website and can be found
 - a. On project website: <https://pinformatics.org/ppirl/faq/faq.htm>
 - b. On GitHub: <https://github.com/pinformatics/mindfirl/blob/master/docs/faq.zip>
- (2) an Institutional Review Board (IRB) application template (attached in the Appendix)
 - a. On project website: https://pinformatics.org/ppirl/faq/irb_app_template_mindfirl.pdf
 - b. On GitHub:
https://github.com/pinformatics/mindfirl/blob/master/docs/irb_app_template_mindfirl.docx
- (3) a template DUA (attached in the Appendix)
 - a. On project website: https://pinformatics.org/ppirl/faq/mindfirl_DUA.pdf
 - b. On GitHub:
https://github.com/pinformatics/mindfirl/blob/master/docs/mindfirl_DUA.docx

Videos

- Creaky Joints webinar presentation, Apr 19, 2018: **Patient Health Data and RL**
 - <https://creakyjoints.org/education/patient-health-data-record-linkage/>
- MiNDFIRL Tutorial Video:
 - https://www.youtube.com/watch?v=xM_Yw4h6nn4&t=12s
- Paper Talk on Kum, H.-C., Ragan, E., Ilangovan, G., Ramezani, M., Li, Q., and Schmit, C. **Enhancing Privacy through an Interactive On-demand Incremental Information**

Disclosure Interface: Applying Privacy-by-Design to Record Linkage. 2019 the Symposium on Usable Privacy and Security (SOUPS). 23% (=27/119 acceptance rate):

- <https://www.usenix.org/conference/soups2019/presentation/kum>
- Paper Talk on Ragan, E., Kum, H.-C., Ilangoan, G., and Wang, H. (2018). **Balancing Privacy and Information Disclosure in Interactive Record Linkage with Visual Masking.** Proceedings of the SIGCHI conference on Human factors in computing systems. ACM. CHI2018 Honourable Mention Award (top 5% of all submissions)
 - <https://www.youtube.com/watch?v=86bE6kOv5sM&t=14s>

Hands on Tutorial

- Short Hands on Tutorial
 - <http://mindfil4.herokuapp.com/introduction>
- Quick Look at the Dynamic On demand Interface
 - <https://ppirl2.herokuapp.com/>

Copyright ©2022 Texas A&M University. All Rights Reserved.

Disclaimer:

The [views, statements, opinions] presented in this report are solely the responsibility of the author(s) and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute® (PCORI®), its Board of Governors or Methodology Committee.

Acknowledgment:

*Research reported in this report was funded through a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-1602-34486). Further information available at:
<https://www.pcori.org/research-results/2016/developing-and-testing-software-linking-patient-data-multiple-sources>*