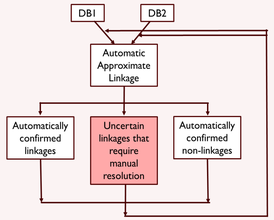
Patients' and Stakeholders' Perceptions of Risk and Benefits of the Privacy Preserving Interactive Record Linkage (PPIRL) Framework

Theodoros Giannouchos[1,4] Hye-Chung Kum[1,2,4] Alva O. Ferdinand[1,4] Cason Schmit[1,4] Gurudev Ilangovan[2,4] Eric D. Ragan[2,3,4]

[1] Department of Health Policy & Management, [2] Department of Computer Science & Engineering, [3] Department of Visualization, [4] Population Informatics Lab (https://pinformatics.org/) Texas A&M University

What is PPIRL ?

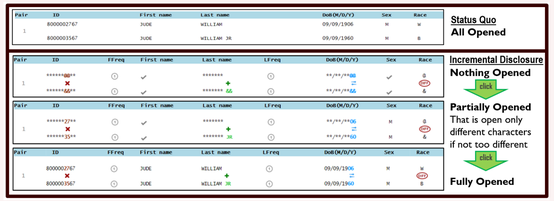
- **Record linkage** is the process of connecting records that belong to the same real-world person in heterogeneous databases
- **Privacy and Record linkage** The absence of a common identifier across databases requires access to personally identifiable information (PII) to accurately link data, which raises privacy concerns
- **Privacy Preserving Interactive Record Linkage (PPIRL)** is a novel approach to enhancing privacy when humans are interacting with PII for record linkage



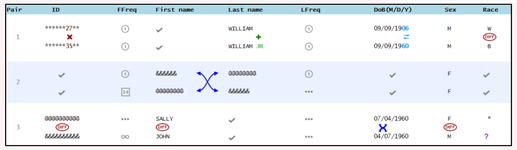
- Typically only 75%-80% can be linked automatically leaving 15%-20% for manual resolution
- Human Interaction with PII is required to produce high quality linked data for data standardization and cleaning, building training data, and tuning model parameters.

PPIRL Features

- **Use of pseudonyms** to separate sensitive information from PII
- **Use of markup highlighting differences** to facilitate decision
- **Minimum Necessary Disclosure**: Hide information that are not necessary
- **Incremental On Demand Disclosure**: Enable partial disclosure of information on an 'as needed' basis (i.e., with a click)
- **Accountability & Transparency**: Measure how much information was disclosed and to whom



PPIRL Framework



Research Objectives

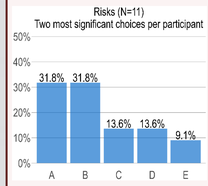
- This research aims to identify patients' and stakeholders' perceptions of the risks and benefits associated with the use of the Privacy Preserving Interactive Record Linkage (PPIRL) framework for record linkage between heterogeneous databases.
- Identifying the perceived risks and benefits of PPIRL will facilitate the design of software that effectively balances conflicting values of information privacy and utility.
- This study will also inform strategies for effective communication with the general public and the institutional review board (IRB) community on the human subject protection aspects of the framework

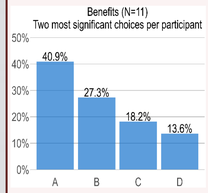
Study Design & Population Studied (N=38)

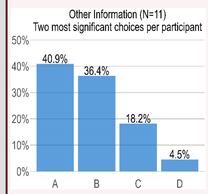
- Six structured nominal group technique (NGT) sessions were conducted to identify perceived risks and benefits of the PPIRL framework: two with IRB and ELSI experts (N=11) and four with patients and caregivers (N=27)
- Participants were given a short introduction to the PPIRL framework and then interacted with an online record linkage and PPIRL tutorial module to enable experiential self-learning
- Attitudes, opinions, and lingering concerns related to the PPIRL framework and human subjects were then explored
- Theme development, discussion, and synthesis followed
- In the final NGT phase, participants voted on what they considered to be the most important themes

Expert Focus Groups (N=11)

Who? ELSI roles: (1) IRB Administrator (program director), Board Member, Manager, Staff, Member, (2) Compliance and (3) ELSI researchers; Organizations: Academic, government, VA, hospital







Q1: What do you perceive as the benefits of using the PPIRL framework for database record linkage?

A.Potential to facilitate the execution of "research protocols" (e.g., providing a tool for researchers to link data, de-identify data, and re-identify 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"

Q2: What do you perceive as the risks for subjects of data when using the PPIRL framework for database record linkage?

A.Potential that software will enable flawed research (e.g., linking flawed data or enabling research that uses inaccurately linked data from user or software errors)

B.Possibility that an organization's administrative controls (i.e. training and user rules) permit inappropriate use of the software.

C.Potential for unnecessary privacy loss/identity exposure to authorized personnel (i.e., among researchers)

D.Potential for privacy loss to unauthorized personnel (e.g. hacking)

E.Potential for a lack of accountability for disclosures

Q3: For research using the PPIRL 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?

A.Evidence for the validity of record linkage when using the software

B.The administrative controls (e.g. organizational rules, policies and required training) and data governance

C.The nature of software security and vulnerability issues, if any

D.Specific details regarding the nature of the data used for record linkage

Conclusion

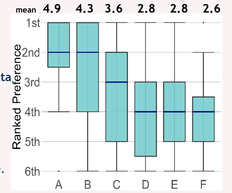
- Enhancing and communicating privacy protection can eliminate existing barriers in the execution of research protocols and can enhance transparency and public trust in the scientific community
- Future work will aim to build consensus on appropriate language for communicating information about the PPIRL framework in patient voice and will inform record linkage software development efforts

Principal Findings

Patient Focus Groups (N=27)

Gender	Male	10	37%
	Female	17	63%
Race	White Non-Hispanic	20	74%
	Black Non-Hispanic	2	8%
	Hispanic/Latino	2	8%
	Asian/Pacific Islander Non-Hisp	3	11%
Income	Less than $25,000	4	15%
	$25,000-$75,000	16	59%
	$75,000-$125,000	4	15%
	More than $125,000	3	11%
Age	Mean=48 (s.d.=15.54)	Min=2	Max=7

Education	HS Graduate or Equivalent	2	8%
	Some College	3	11%
	College Graduate	15	56%
	More than College	7	16%
Type of Health Insurance	Medicare	4	15%
	Medicaid	1	4%
	Dual	3	11%
	Commercial	16	59%
	Va or DoD	1	4%
	Other	2	8%
Years w/ Disease	Mean=14.2 (s.d=13.77)	Min=1	Max=5



Q1: Are there things you like about the software that you would tell your neighbors? *Ranking (1=least important, 6=most important)*

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

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

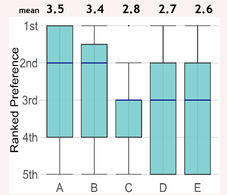
C.Software allows for participants to feel good about the use of their data in a safe manner while still having confidence in the quality of the results.

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

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

F.Software allows for tracking disclosures to enhance accountability

Friedman = 31.88 (P-value < 0.0001) , Post hoc Nemenyi test: (1,2) 3 (4, 5, 6)



Q2: Are there concerning things about the software that you would tell your neighbors? *Ranking (1=least important, 5=most important)*

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

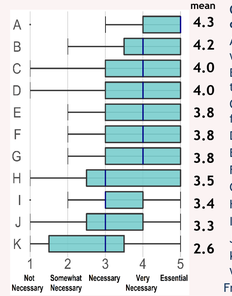
B.Still potential for misuse of information by authorized users (e.g. negligence, not sufficient training)

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

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

E.Still potential for errors in the linkage process

Friedman = 8.18 (P-value = 0.085)



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

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

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

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

D.Why are identifiers needed for this research?

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

F.Where can I get more information?

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

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

I.What other information, besides personal identifiers, are used during linkage?

J.How will the results be disseminated?

K.Has the software been used before for research and has it enhanced protection as well as improve research quality?

Friedman = 34.33 (P-value < 0.0001)

Research supported in part by the Patient Centered Outcomes Research Institute ME-1602-34486