Enhancing Privacy through an Interactive On-demand Incremental Information Disclosure Interface: Applying Privacy-by-Design to Record Linkage

Hye-Chung Kum[1], Eric D. Ragan[2], Gurudev Ilangovan[1], Mahin Ramezani[1], Qinbo Li[1], Cason Schmit[1]

[1]*Population Informatics Lab, Texas A&M University*   [2]*INDIE Lab, University of Florida*
{kum, ilan50_guru, mahin, lee, schmit}@tamu.edu;   eragan@ufl.edu

Abstract

Achieving the benefits of data science in cases involving personal data requires the use of that data, which results in some privacy risk. Our research investigates approaches to enhance privacy while supporting legitimate access for human decision making by capitalizing on the fact that in most human-computer hybrid systems, only a small fraction of the full data is required for human judgment. We present an interactive visual system for record linkage – a task that requires human decision-making about whether different but similar data records refer to the same person. The system employs an on-demand interactive interface that incrementally discloses partial information only when needed and other feedback mechanisms to promote ethical behavior. We evaluate our approach with a controlled experiment of how different types of feedback and access restrictions affect human decision-making quality, speed, and access behavior. The on-demand interactive interface reduced privacy risk to only 7.85%, compared to 100% when all data is disclosed, with little to no impact on decision quality or completion time. In addition, feedback from an expert review supports the notion that an intermediate level of access other than "all or nothing" can provide better accuracy than no access but more protection than full access.

1. Introduction

The potential impact of population informatics—data intensive secondary analysis of large integrated population data—are endless [1]. Access to such data for qualified researchers could provide a greater understanding of root causes of social and public health problems, help identify upstream opportunities for interventions, help predict the downstream effects of different policy options, and assist in allocating our collective resources for the greatest impact to benefit our society. As one example, a National Institute on Drug Abuse (NIDA) study integrated data from multiple databases on 56,923 Medicaid beneficiaries with opioid dependency to conclude that buprenorphine was cheaper and safer than alternative treatments [2, 3]. Another example is a three state study that integrated three data systems to follow children from the foster care system for over 10 years to investigate long-term employment and income trends [4].

Figure 1: Pairs of PII for human judgment in record linkage

While the potential benefits of population informatics are clear, access to such population data for these research is not widespread and is often given on a one-time basis for a single project. In fact, when compared to the widespread use of population data in other less regulated sectors such as marketing, intelligence, and campaigning, secondary analysis of population data for research is quite restricted and lacks infrastructure. This is especially true in the United States, where privacy concerns make it difficult to build and maintain large integrated population data for research. In contrast, Canada, UK, and Australia have invested in establishing population data linkage centers [5].

One of the core challenges in establishing integrated population databases is addressing privacy concerns during data integration. High-quality data integration requires record linkage (RL), the process of identifying records from heterogeneous data sources that potentially refer to the same person in cases where a common identifier is not available. Privacy becomes a major issue because one must exactly identify the identity of records to accurately build the integrated data. During RL, it is important to distinguish between family members or twins [6] and to handle changes in the data (e.g., change of last name) and data errors. Thus, most projects that require integrated data obtain access to *personally identifiable information* (PII) (e.g., names, birth dates) to use for RL. Figure 1 shows a simple example of different types of data discrepancies in PII pairs.

In practice, most linkage projects use semi-automated linkage systems where the majority of the linkages are made using algorithms that humans have to tune, maintain, and manually resolve more complex cases. Properly using auto-

mated methods requires a significant level of human involvement for data cleaning [47], standardization, training data construction, parameter tuning [46], and validation [6, 7]. For example, Bronstein et al. [8] describe the process of matching pregnancies from Medicaid data to birth records using probabilistic record linkage that involved 11 manual steps. The process required many human decisions in the process such as cleaning up 4,369 pregnancies linked to more than one vital records. In another study linking cancer registry data to health service data, 15% (16,288 links) were confirmed through manual verification [9]. Such a high level of human interaction and iteration is common in medical record linkage studies [8-10]. Without these human involvement, the match rate and biases in these studies would be problematic because the errors in the linkage step propagate downstream to all analysis using the integrated data. In addition, automated linkage can also result in selection bias such as in preferentially selecting patients with complete information on required identifiers, which can under-represent particular groups such as socioeconomically disadvantaged and racial/ethnic minorities [5, 8, 11, 12, 13].

Human review often involves looking at similar pairs of records to make judgments about complex cases where different records could potentially refer to the same person or to decide how to tune linkage algorithms to improve data integration. These tasks require disclosure of some PII to some people, which poses potential privacy risks. In most linkage projects, this usually means that staff get full access to all PII in a dataset to enable quality linkage decision-making [5, 8, 9]. Naturally, such unlimited access to PII raises concerns about privacy and has led to many efforts to develop automated privacy preserving RL algorithms. Most algorithms securely compute a known linkage function using encryption and trusted third-party computing. These approaches assume *no access to PII as a solution to privacy*. However, such solutions are problematic for assuring quality data integration for complex cases requiring human judgment, especially since the details of the linkage function is rarely known ahead of time. Thus, open questions remain about (1) how to determine the linkage function, (2) support the human tasks required to obtain high-quality linkages, and (3) how to validate the linkages found [14, 15].

Our work on privacy-enhanced RL takes a fundamentally different approach to privacy in order *to produce high-quality validated results by enabling human judgment where needed*. Rather than rely on various security technology to limit access for privacy, our premise is that human access to some PII is necessary and encouraged for valid results. Hence, we rely on two fundamental principles of privacy to promote legitimate and confidential access: (1) *the minimum necessary* principle and (2) *accountability through transparency* principle. Thus, the critical questions for privacy enhanced system design are: (1) What and how much infor-

mation about the PII needs to be accessed for good linkage decisions? (2) When do you know what you need to access? (3) What accountability mechanisms would be effective to discourage bad behavior? Our approach is motivated by the hypothesis that there is some level of partial disclosure of PII—between unrestricted access to all PII and no access to PII—that can effectively support human judgment and validation while significantly reducing total PII access.

In this paper, we present and evaluate a novel privacy-enhanced RL system (see Figure 2) for safe human interaction with PII. The core tenet of our method is an on-demand interactive method for incrementally disclosing limited information, only as-needed, and when explicitly requested. This approach makes it possible to meet the legal requirements for minimum necessary information disclosure standards while also enabling accountability through the ability to log access requests to individual PII details. The contributions of the presented research are threefold:

- First, we present our interactive record-linkage system that uses (1) on-demand, incremental information disclosure, (2) feedback of privacy risk, and (3) enforcement of disclosure budgets to facilitate high-quality decision-making while limiting overall access to personal data.
- Second, we present a controlled experiment to evaluate how different types of feedback and access restrictions affect human decision-making quality, speed, and access behavior in a record linkage task.
- Third, we also present an expert review with domain scientists who regularly conduct research with PII.

2. Background and Related Work

In this section, we provide the basis of our approach in the privacy regulations and review relevant privacy literature.

2.1. Minimum Necessary Standard and Practical Challenges

As discussed in detail in the introduction, human review of personal data is common for a variety of data work and required for valid results [6-10, 46, 47]. Research on information privacy has shown the complex nature of providing protection while still allowing utility from legitimate use of personal data for social benefit [16]. 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 many different data protection laws in the form of *minimum necessary* or *need-to-know* information disclosure standards. Laws like the *Health Insurance Portability and Accountability Act* (HIPAA), the *Privacy Act of 1974*, and the confidentiality protections for substance abuse disorder records in *42 CFR Part 2* use similar legal standards to permit legitimate uses of data while protecting privacy by limiting extraneous disclosures [50-52]. Similarly, the EU General Data Protection Regulation (GDPR), uses the principle of "data

minimisation" to limit data use to what is necessary for a permitted purpose [53].

However, practically implementing a process for sharing protected data that restricts disclosures to the minimum necessary is a daunting task [54]. It is rarely the case that a data project knows exactly what 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 changes over the course of the project. Practically, in many situations, it is the case that *all* the data is decided to be the "minimum necessary" [55].

These dynamics can lead to serious consequences when negotiations and legal agreements must be made (e.g., data use agreements) between different organizations for data sharing. Perceptions about what constitutes the minimum necessary can differ between data sharing partners, leading to prolonged project delays [56]. Even worse, funded projects may be cancelled when researchers are not able to pass a vetting process for giving full access to protected data [56]. One reason for this is because there are no practical tools to facilitate data disclosures that better meet minimum necessary legal standards. Our research addresses this need.

2.2. Privacy and Human Behavior

Researchers have explored a variety of approaches to system design [17] and interface design to support privacy enhancements [18]. For example, Iachello et al. [17] describe the design process for a privacy aware social location disclosure application through a series of user studies. They present a list of privacy guidelines from these studies demonstrating the privacy by design approach. Dasgupta et al. [19] presents metrics for privacy as applied to visualization. They demonstrate the use of aggregation, clustering, and uncertainty in scatterplots and parallel coordinate plots to allow inspection of sensitive data while limiting knowledge of individual elements. In contrast, our work focuses on data inspection tasks that require review of individual PII for accurate decision making.

As an example from our prior work involving access to individual PII, Ragan et al. [20] demonstrated how the use of visual masking techniques could be used to hide data values in tabular data interface while still showing differences to support data cleaning and de-duplication tasks. Kum et al. [6] also studied different mechanisms (i.e., deception, obfuscation, and blurring about the nature of the list of names) to hinder inference of identity when names are disclosed. They found that these methods were effective in introducing uncertainty to protect the real identities of names for both common and rare names. Work by Hasan et al. [21] used a similar approach but for images. The authors studied visual

obfuscation methods for hiding or altering portions of photographs to preserve privacy, and they discuss the tradeoffs of different approaches in terms of both privacy and the effects on the general interference or distraction when viewing images. Similarly, Çiftçi et al. [22] demonstrated how altering the color composition for facial images can make it difficult to recognize people in photographs.

Prior research has also demonstrated that users' behavior or attention to privacy can be influenced by their experiences with technical systems. For example, Chang et al. [23] found that participants' inclination to disclose information could be influenced by the types of profile pictures they observed prior to the interaction. When viewing less revealing profile images, the participants were less likely to share their own personal information. These results suggest that decisions for acceptable privacy behavior might be influenced by the perception of what others find acceptable. John et al. [24] ran similar experiments asking participants whether they had engaged in a number of sensitive activity (e.g., sexual behaviors). They measured the proportion of questions answered affirmatively as an indicator for privacy concerns and varied the look and feel of the website (i.e., professional, baseline, unprofessional). Those who were asked on the unprofessional website were almost twice as likely to admit to engaging in the sensitive activities compared to the baseline and professional websites indicating that disclosure of private information responds to environmental cues. The results support the general idea that a system may influence a user's attention to privacy by using different cues.

In fact, a comprehensive review of multi-disciplinary literature presents multiple interventions (e.g., education, feedback, framing, positive and negative incentives) that can be used to influence privacy decision making [25]. In this research, the main intervention tested is feedback. Password meters is a good example of how effective feedback systems can nudge to create stronger passwords [26]. Our research studies a feedback mechanism similar to the password meters that gives real-time feedback and allows decisions to be altered based on the feedback given.

2.3. Quantifying Information Privacy

Quantifying the privacy risk is an active area of research with the best approach being context dependent [27]. k-anonymity was the first method proposed based on the insight that a record may not be distinguished from at least k-1 records when there are k shared records [28-31]. Machanavajjhala et. al. [32] have shown issues with k-anonymity when there is a lack of diversity allowing for background attacks and introduced the l-diversity model that aims to have intra-group diversity for sensitive values. Li et al. [33] have shown that (1) l-diversity may be difficult and unnecessary to achieve and (2) l-diversity is insufficient to prevent attribute disclosure. To address these problems Li presented
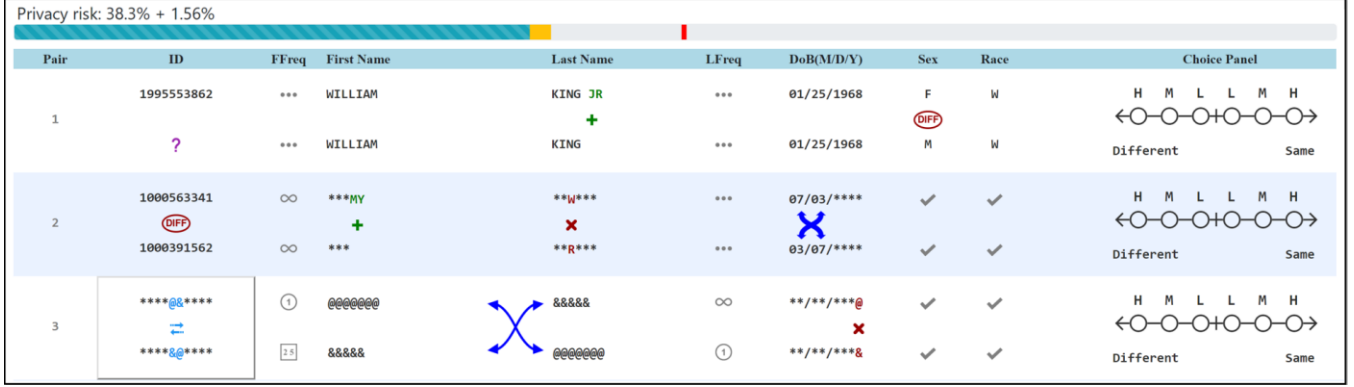
Privacy risk: 38.3% + 1.56%

Pair	ID	FFreq	First Name	Last Name	LFreq	DoB(M/D/Y)	Sex	Race	Choice Panel
1	1995553862	•••	WILLIAM	KING JR	•••	01/25/1968	F (DIFF)	W	H M L L M H ←O–O–O+O–O–O→
	?		WILLIAM (+)	KING		01/25/1968	M	W	Different Same
2	1000563341 (DIFF)	∞	***MY (+)	**W*** (✗)	•••	07/03/**** (✕ swap)	✓	✓	H M L L M H ←O–O–O+O–O–O→
	1000391562	∞	***	**R***		03/07/****	✓	✓	Different Same
3	****@&**** (⇄)	①	@@@@@@@	&&&&&	∞	**/**/***@ (✗)	✓	✓	H M L L M H ←O–O–O+O–O–O→
	****&@****	25	&&&&&	@@@@@@@	①	**/**/***&	✓	✓	Different Same

Figure 2: Example from the study application showing (1) supplemental markup and value masking, (2) interactive clickable interface, (3) feedback privacy meter, and (4) the privacy budget (solid red line on the meter). The visual markup highlights discrepancies, provides information about name frequency, and hides common values. The box on the last row indicates that the user has moused over this area, and is considering whether to click open or not. The user should be taking into account the feedback meter on top which indicates the accumulative disclosures to now in blue, and what additional risk will occur if the selected information is clicked open in orange. Finally, the solid red line on the meter indicates a limit to the disclosure that the user can request.

Highlight discrepancies

? Missing fields
✗ Different characters
✚ Extra characters
⇄ Transposed characters
✕ Name/date swaps
(DIFF) Major field differences

Highlight data details for privacy

✔ Same fields
*** Same characters

Name frequency meta-data

① Unique
25 Rare
••• Common
∞ Highly common

Figure 3: Visual masking icons used to highlight discrepancies, including matching values, and providing meta-data [30].

t-closeness based on the differences in the distribution of the sensitive attribute [33]. In another study, Li et al. [34] presented another approach based on k-anonymity and differential privacy and used a method of input perturbation to add uncertainty. Currently, differential privacy models provide the strongest guarantees and is the most active area of research [27, 35]. In particular, many differential privacy algorithms have been proposed to answer low dimensional counting queries. However, adoption of these methods in practice has been limited due in part to the wide variation in error rates, which are dependent on the properties of the input data [27]. It is important to note that although these approaches are related and may be applicable to the work in this paper, quantifying the identification risk to support user decisions to disclose certain parts of the PII in RL is different from risk in low dimensional counting queries. Although k-anonymity does not address sensitive attribute disclosure, it is a well-established measure for identity disclosure, the focus of this work [36], and our paper presents our first approach based on k-anonymity with the incorporation of differential privacy in progress.

3. System Design

Our research contributes a novel interactive interface where we start with fully-masked de-identified 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 [15] 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 2). The system takes advantage of three 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 data access via a budget. Before we present our study of how these techniques can affect privacy while still maintaining the quality of the linkages made, this section describes the system in terms of its core mechanisms and design rationale.

3.1. Design Rationale

Our research capitalizes on the fact that in most human-computer hybrid systems for sensitive data, only a tiny fraction of the full data is required for tasks requiring human judgement. Prior research provides evidence to suggest that the optimal level of disclosure to achieve high quality linkage is quite low with minimal risk to identification when appropriate meta-data is shared using data masks [20]. Ragan et al. investigated the effectiveness of different levels of disclosure on static interfaces [20]. While the results are promising, the tested system only supported static, pre-specified levels of data hiding.

Our research investigates dynamic *just-in-time incremental* techniques to enhance privacy in data systems requiring human access to personal data for legitimate purposes. We present an interactive visual system (see Figure 2) for linking personal data using an on-demand interface that incrementally discloses limited information—and only when

needed and explicitly requested. This approach minimizes data disclosure to optimal levels for human judgment while observing legal requirements to follow a *minimum necessary disclosure* standard. Further, the system's interactive interface satisfies accountability requirements since all disclosure occurs via explicit user actions, which makes it trivial to log who accessed what data.

In addition, we study different design mechanisms to promote accountable ethical behavior in information access decisions. The system incorporates visual feedback and allows access limitations to encourage conscientious data review. By quantifying and displaying the privacy risk associated with each increment of disclosure, the system can help users to consider the tradeoffs between privacy and decision quality for each piece of information. Further, an optional maximum *disclosure budget* can be enforced to provide guidance to novice users about the right balancing point for good decision making, or to meet external requirements for accessing sensitive data.
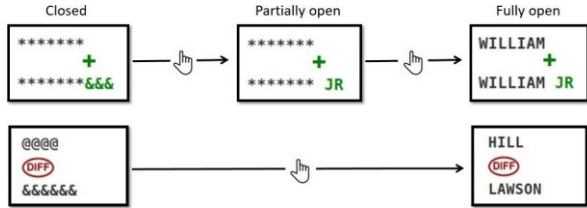
Figure 4: Interactive on-demand interface. Cells start with no disclosure, then partially open with a click. Cells open fully with either 1 or 2 clicks depending on the nature of the data.

3.2. Minimum Disclosure via Interactive Just-in-Time Interface

The system is designed to support minimum disclosure for interactive RL using dynamic information access. When linking records, the reviewer's task is to consider the discrepancies and make a decision whether the two records corresponds to the same person or different people. A choice panel on the right side allows fast recording of each decision (shown on the right-most side of Figure 2), and the interface can also record the degree of user confidence by high, moderate, and low levels (denoted by *H*, *M*, and *L* labels).

The first part of enabling the on-demand design is to have the default state of the interface hiding all characters and instead use visual icons and meta-data to help highlight how the records in each pair differ. Figure 3 shows an overview of the primary icon types used for visual masking [20]. The bottom row of Figure 2 (*Pair #3*) shows an example of this default masked state. As seen in the "Sex" and "Race" fields, checkmark icons indicate cases where the field contents are exactly the same. Otherwise, asterisks and replacement characters summarize differences, and supplemental icons help explain the type of differences. Additionally, icons indicate frequency of both the first name and the last name (by the *FFreq* and *LFreq* columns, respectively) in each data source since linkage decision-making can depend

on how rare or common names are in the source data. A more detailed explanation of the visual masking techniques and an evaluation of their effectiveness for decision making are presented in [20].

The focus of this paper is the creation and evaluation of a new method for accessing details only as needed. As seen in Figure 4, the user is first presented with only the de-identified data using visual masks. If users need more information to make good linkage decisions, they can click on the cell (i.e., a specific field for each PII pair) to reveal more details. The first click will disclose only the characters that are different between the cells for that pair, and an additional click on that cell will show the full cell contents. Depending on the nature of the differences in the pair, cells will open fully with one or two clicks. While users have the choice to access different levels of detail, the example pair in the middle row of Figure 2 (*Pair #2*) shows examples of partial disclosure, where some characters are visible to aid inspection of differences. The top row (*Pair #1*) shows full disclosure without character masks, which would be possible if the user fully clicked each individual cell in the row. By enabling different levels of disclosure that only reveal data with an explicit click, the system effectively supports the *minimum necessary* principle by preventing most of the PII from being accessed unless necessary for good decisions.

3.3. Accountability via Quantified Privacy Risk

While the interactive method for just-in-time incremental disclosure can support minimum necessary access to PII, it allows users to decide what information is accessed. To promote preferred privacy-aware behaviors and prevent misuse of sensitive data, the interactive interface is augmented with mechanisms for information accountability (i.e., transparency and continued monitoring) and audits for misuse [37]. Concretely quantifying the privacy risk and making this information available to everyone (e.g., data workers, managers, and compliance administration) is the start of accountable access to PII.

Quantifying the privacy risk and providing feedback on it also supports good human decisions because decisions can be improved when it is informed by relevant information [26]. To impact the decision, the information must be concrete enough to be actionable. Generally, instructing researchers working on linkage to "disclose as little as possible" is not actionable. But, when the actual risk of identification from disclosing a piece of data can be concretely quantified and shared ahead of time to inform the decisions to view the data, we believe that people will be encouraged to make more thoughtful decisions based on the risk level.

To this end, our system uses a method for quantifying privacy risk based on factors such as amount of characters accessed, type of information, and its uniqueness; then, the interface uses this privacy measure to display feedback

about the risk associated with each data-access decision. Feedback is shown by a visual meter (see the top of Figure 2), where the length of the meter represents 100% disclosure, the blue bar represents the current accumulative access, and the temporary orange extension represents the added risk of disclosure for the currently selected cell that the user has moused over. If the user decides to click on a masked cell, then the data will be shown, the privacy cost will be used, and the meter will update to show a new level of disclosure. If the user moves the mouse off of the cell without clicking, the "hypothetical" orange increase to the meter bar goes away. The recording of these clicks and the feedback have similar role to how a surveillance camera can encourage good behavior.

A number of factors were considered for risk quantification. Measuring the identity disclosure risk for a given partial disclosure of personal data is not trivial because not all pieces of information lead to the same level of identification. Mathematically, the identity disclosure risk is inversely related to the number of entities in the population that share the information disclosed. If the information refers to one and only one person in the population, then the uniqueness of a person's identity information makes it easy to match the information to a specific person. On the other hand, if the disclosed information is identical for multiple people, then the information is less revealing, as it could refer to any one of those people. Quantifying privacy risk is an active area of research with the best approach being context dependent [27].

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 identifiers. Thus, our prototype used the k-Anonymity Privacy Risk (KAPR) score which uses the anonymity-set size as an estimate of the identity disclosure risk. *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 (e.g., Eric) is disclosed, there is a low probability that a specific person with that name could be identified. In comparison, a rare name (e.g., 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.

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. Uniqueness is calculated based on the data being linked. An example and the exact measure can be found in [38]. 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 feedback meter that users can understand. Further research is needed to study the trade off between using easy to understand functions (e.g., percentage of information disclosed) versus more accurate but complex function (e.g., KAPR score) for quantifying the privacy risk.

3.4. Limiting Privacy Risk via Budget

Although the interactive interface enables only the minimum necessary disclosure and the feedback meter encourages limited access behavior through accountable access to PII and audits after the fact, neither of these designs alone can enforce limited disclosure that may be a condition of use. For example, certain data usage agreements may limit access to social security numbers by allowing up to four digits. In our system, such hard rules on data access can be enforced using an option to configure the interactive interface with hard rules ahead of time. In particular, the privacy budget feature can be used to enforce a limit on the total disclosure for a given use case.

By specifying an allowable privacy budget ahead of time, the system can guarantee a certain level of information disclosure. 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 versus trying to make do with limited access to 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 research. 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 record linkage task.

3.5. Threat model

The main threat model for this work is the insider threat model where the system goals are to minimize any incidental knowledge from legitimate access to PII, and discourage against access for unauthorized purposes by authorized users. First, the on-demand interface will minimize any incidental privacy risk of data workers seeing information about people they know (e.g., co-workers). In addition, quantifying the privacy risk with the meter feedback discourages insiders from abusing their ability to access information. This is similar to surveillance cameras that

discourage people from bad behaviour by making it possible to enforce accountability. To operate the system effectively, having clear reporting and audit processes in place for the logs will be important just as with camera footage. Although cameras cannot gurantee no bad actors, it is very effective in keeping people on good behaviour, especially when it is clearly displayed. Finally, enforcing a limited budget provides further ability for mangers to manage risk from insiders at acceptable levels. Managers may set low limits on disclosure ahead of time and iteratively increase the limit as requested when the context requires high levels of privacy.

4. Experiment

Using our interactive record-linkage system, we conducted a controlled experiment to evaluate how different mechanisms for privacy protection affect information access and decision-making for tasks requiring interpretation of PII.

4.1. Hypotheses

Our over-arching 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 three 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 pre-specified budget on the meter. Our evaluation of these mechanisms follow three 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. The quantification of the risk and visibility of this information for all relevant parties (e.g., 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 pre-specified 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, then 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 in [20], 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.

4.2. Experiment Design

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

- *Fully open*: non-clickable 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 meta-data, but no data was 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 2) 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 pre-specified limit on the privacy budget indicated by a thick red line on the meter. This condition sets the limit at a moderate disclosure level believed to be sufficient to make good linkage decisions. The specific limit in this condition was 35.7% to 37.8% KAPR score depending on the specific dataset. This amount was chosen based on the *moderate* level from a prior study [20] that focused on static, non-interactive levels of information disclosure. 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 as in Figure 2. This condition sets a lower limit with KAPR scores ranging from 5.02% to 6.48% depending on the dataset. This level was again chosen based on a previous study [20], 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 with-

out compromising linkage decisions when interactive disclosure is used.

Figure 5 shows a simplified summary of the differences among the five conditions. The conditions allow us to test our hypotheses about the effects of different mechanisms to discourage unnecessary disclosure. We address hypothesis H1 by comparing the results from the *fully open* to the *no meter* to determine how much more we can reduce disclosure using the interactive interface. Hypothesis H2 compares the *no meter* to the *unlimited meter* to determine if a feedback meter is effective in reducing unnecessary disclosure. Finally, hypothesis H3 compares 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 linkage.
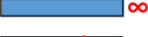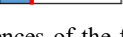
Condition	Default Masking	On-demand Interface	Meter & Limit
Fully open			
No meter			
Unlimited meter			
High Limit			
Low limit			

Figure 5: Visual summary representing the differences of the five experimental conditions in the evaluation.

4.3. Generation of Test Data

To allow us to evaluate the effects of the different system configurations on record linkage performance, we had to have data pairs that could serve as "ground truth". Since real scenarios do not have known "true" answers, our experiment used a derived data set created by modifying publicly available voter registry data (as in a previous study [20]). The generated test data comprised of realistic pairs of records based on a large county's records from 2013 and 2017. To establish a known ground truth, the registry number and address information were used to identify the same people among many generated pairs in the original data. We also tweaked the pairs to control for the kinds of differences and emulated real world data errors like typographical errors, family relationships (e.g., twins and siblings), name changes, field swaps, and missing fields.

In total, we had 747 pairs of records with "same" or "different" labels. These pairs were used to generate 10 random samples of 36 questions each, and each user was randomly assigned one such sample. It should also be noted that out of the 36 questions, 6 questions (one in each page) were easy questions for which the answers were obvious (for example, all different fields would mean the pair referred to different people and vice versa). These questions primarily served as attention checks and to verify that participants had sufficient understanding of the decision-making process for linkage.

4.4. Procedure

The study was approved by our organization's Institutional Review Board. We note that the procedure for this study was designed to be similar to a previous study using an interactive record linkage activity [20]. The study was run in group sessions in a computer lab, but each participant worked independently. Each study session lasted two hours. The system was run as a web application on Google Chrome. All participants used identical computers running Windows 7 with 23-inch displays at 1920x1080 resolution.

To begin, participants completed a background questionnaire to collect information about age, gender, education, academic specialization, experience with data analysis, and primary language. Next, the experimenter gave participants an overview of record linkage, the system, and the instructions for the task. Participants then worked through the system's tutorial, which included sample questions and additional instructions. To help participants understand the decision making, the tutorial provided the correct answers for any practice linkage questions that were answered incorrectly, and participants had the option to repeat or review all information. Different configurations of the tutorial were designed to match each experimental condition, and the final phase of the tutorial had participants work through 36 practice questions.

After the tutorial, participants started the main linkage trials, which were organized into multiple sets of 36 linkage questions shown in groups of six questions per page. Participants worked through as many sets as they could complete in the study time. Finally, near the end of the study session, participants concluded by completing a closing questionnaire that asked for comments about the linkage task, the system, and their comfort with sharing personal information.

The paper presents data from the first set to 36 linkage questions because everyone completed at least one. Analysis of the second set had comparable results with the first set, but fewer participants. Full analysis scripts and data are publicly available at [48].

4.5. Participants

The study had a total of 122 participants, and each participant completed one condition. We used the 6 trivially easy questions to filter participants who did not demonstrate sufficient effort or competency for the linkage task. Participants who incorrectly answered more than one of the easy questions were excluded from analysis. Two participants failed to meet the requirement, and hence their data was excluded. Thus, data from 120 participants were considered for data analysis. Of these, 22 were in fully open, 23 were in no meter, 26 were in unlimited meter, and high limit, and 23 were in low limit. The final numbers in each group varied due to the competency screening and the study being run in prescheduled lab sessions. 55.8% of the participants were female and 44.2% male. Ages of the participants ranged from

18 to 42 with a median age of 22. Participants came from diverse academic fields. 52.5% of participants were either pursuing a graduate degree or already had one, and the remaining participants were undergraduate students.

5. Experiment Results

We analyzed the results to test for differences based on the previously explained hypotheses. We did not conduct statistical comparisons of all five conditions together because this would confound the presence/absence of different mechanisms. When testing for statistical differences in measure (error rate, KAPR scores, duration), we conducted either a parametric test (Student's t-test or Welch's t-test) or a non-parametric test (Wilcoxon rank sum test) depending on the data and if the assumptions of parametric testing were met. The procedure we followed was to test the measure for normality using the Shapriro-Wilk test. If the measure passed the normality assumptions, we did an F-test to test for the homogeneity of variance. Further, we tried data transformations (e.g., log or square root) to satisfy assumptions when possible. If the measure passed the normality and F-tests, we used a Student's t-test. If it passed the normality test but failed the homogeneity of variance test, we used the Welch's t-test which accounts for different variances. If it did not pass both, we used the non-parametric Wilcoxon test. For all tests, we used a base alpha level of 0.05 and applied Bonferroni correction for the four hypotheses, which resulted in an adjusted significance threshold of 0.0125.

5.1. Performance Overview

Risk of privacy loss was calculated using the KAPR measure which calculates the actual risk of identification (i.e., how unique the revealed information is) based on what information has been disclosed. Across all conditions, the score ranged from 0% to 100%, with overall mean of 23.31% (SD = 36.79). Figure 6 and Figure 7 show the error rate and risk score results broken down by condition. We present the results using violin plots which, in addition to displaying the median and interquartile range, also show the distribution of the data [57].

We also consider completion time (see Figure 8), which includes only the portions of the study spent answering the main 36 record-linkage questions. Analysis of the differences in participant confidence in linkage decisions had similar results to a previous study where confidence was lower for incorrect responses, which suggests their lack of confidence was justified [20].

5.2. H1: Effects of interactive on-demand disclosure

Hypothesis 1 is concerned with differences in information disclosure between the baseline static interface with all information already visible and the on-demand interface starting with no information but incrementally reveals more when participants need to see more. So we compared condi-
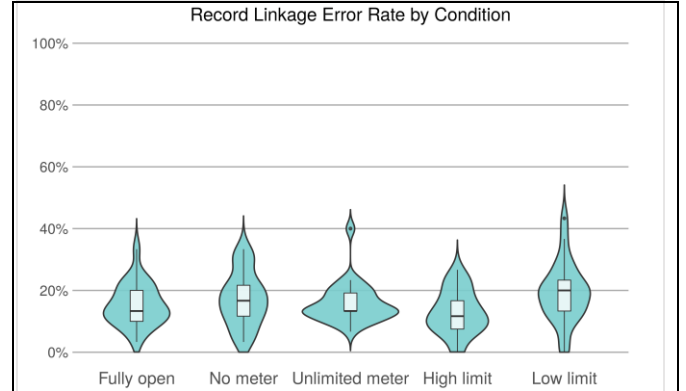

Figure 6: Percent of incorrectly linked pairs from the five conditions. Lower values indicate better performance.
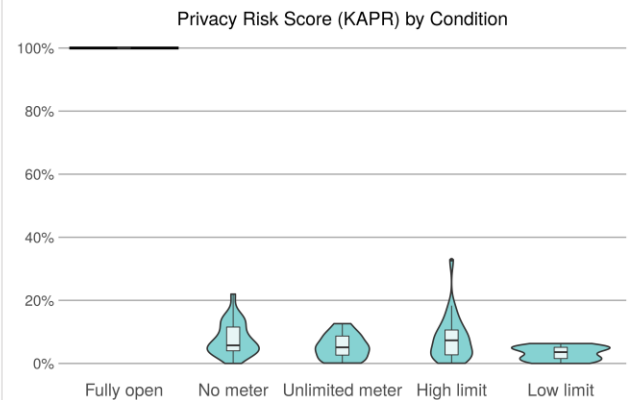

Figure 7: KAPR privacy scores for the five conditions. Lower scores indicate lower risk. Note the *fully open* condition has 100% privacy risk score due to all characters being visible by default.
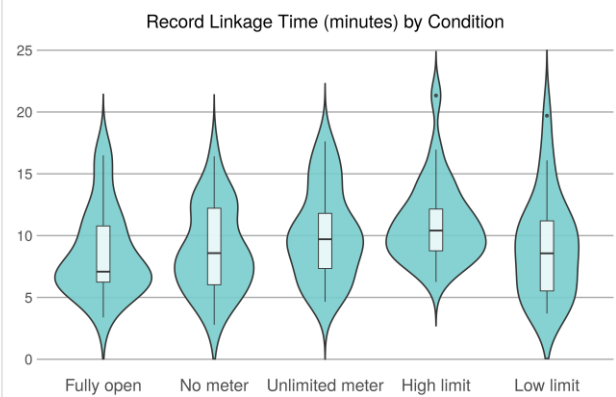
Figure 8: Time taken to complete the linkage task for the five conditions.

tions *fully open* and *no meter*. The *fully open* condition had full data disclosure with no data hiding, so the privacy risk score was a constant 100% KAPR for all participants. In comparison, *no meter* had an average of only KAPR = 7.85% (SD = 5.23), indicating very low levels of disclosure and risk even though the participants could reveal as much information as they wanted. While this is an obvious difference in information disclosure, we can also verify this

through inferential testing. A Wilcoxon rank sum test found a significant difference with Z = 44 and p < 0.001.

Even with such low levels of disclosure in *no meter*, the error rate did not increase significantly when compared to *fully open* (see Figure 6). A student t-test did not find a significant difference between the error rates at t(43) = 0.77. Though no difference was found, we cannot definitively claim that the on-demand disclosure method did not induce an increase in the error rate.

We also tested for differences in completion time between the different modes. A Student's t-test on the log-transformed completion times found no significant difference at t(43) = 0.85. These results support H1, demonstrating that the on-demand design significantly reduced disclosure (to only KAPR = 7.85%) with little impact on decision quality or speed.

5.3. H2: Effects of feedback on the quantified privacy risk

Hypothesis 2 is concerned with the differences in information disclosure when the feedback meter is added to the on-demand clickable interface. Therefore, to address H2, we compared *no meter* and *unlimited meter*. We tested the effects of adding a feedback meter on privacy by performing a Student's t-test on transformed KAPR scores. The tests failed to detect a significant effect with t(47) = -1.83 and p = 0.07. We note that the risk score was lower when the meter was added (see Figure 7), and more data might lead to statistical differences, but further experimentation would be needed. In addition, it is worth noting that the *no meter* condition already had very low levels of disclosure leaving little room for improvement. Regardless, a Wilcoxon test on the error rates found no significant difference with Z = 327. A Student's t-test on the completion times also did not find a significant difference at t(47) = -1.23.

Thus, the study results were unable to provide evidence for H2. Both quality of decision and completion time were similar, and although adding the feedback meter to the interactive on-demand disclosure reduced the KAPR score from 7.85% to 5.33%, this difference was not statistically significant. However, the relatively low p-value (p = 0.07) suggests the results may be inconclusive and motivates further study, especially considering other findings indicating that people may change privacy behavior with appropriate feedback which is consistent with literature [25, 26].

5.4. H3: Effects of limiting the privacy budget

Hypothesis 3 is concerned with differences in information disclosure and linkage decisions given different limits in privacy budgets. First, we compare *unlimited meter* and *high limit* to test *H3.1*, then we compare *high limit* and *low limit* to test *H3.2*. We do not compare all three conditions together because *low limit* would affect the quality score, confounding the relationship between limit and disclosure.

Hypothesis *H3.1* did not hold in the comparison between *unlimited meter* and *high limit*. A Student's t-test on transformed KAPR scores failed to detect significant differences with t(50) = 1.46. And a Wilcoxon test failed to detect any difference in error rate at Z = 354.5. Finally, a Student's t-test on log-transformed completion times found no evidence of differences at t(50) = -1.25.

Although disclosure levels were higher in the *high limit* condition (7.87% vs 5.33%) compared to not having a limit, it did not near the specified budget (M = 36.7%, SD = 0.81). On average, participants used only 21.4% (SD = 19.1) of the given budget. Thus, the results do not provide evidence in support of hypothesis *H3*.

Providing a high limit did nudge participants to disclose slightly more in the *high limit* condition. But 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, hypothesis *H3.2* did hold in the comparison between *high limit* and *low limit*. We performed a Welch t-test on the KAPR scores, which showed evidence of differences in the risk scores with t(35.8) = -3.46 and p<0.001. For participants given the *low limit* condition, KAPR was less than half (M = 3.22%, SD = 2.12) compared to those given the *high limit* condition (M = 7.87%, SD = 7.09). Although this accounted for much more of the given limit (M = 57.6%, SD=36.4) compared to the high-limit condition (M=21.4%, SD=19.1), it was still much less than the given budget.

A Student t-test on the error rate scores found a significant difference between the modes at t(47) = 2.62 and p = 0.012. A Welch t-test on the log transformed completion times found that there was also no significant difference in completion times between the modes (t(36.21) = 2.3 and p = 0.027) at the Bonferonni-adjusted α = 0.0125. The error results indicate that the quality of human decisions will suffer if low disclosure limits are enforced.

In sum, the interactive on-demand interface was effective in reducing disclosure to very low levels while still supporting good decisions. In addition, there is some evidence that feedback using the risk quantification may further discourage unnecessary access to PII. Limiting access via a pre-specified budget may influence disclosure decisions, but more research is needed to design optimal systems to induce best behavior. Finally, the results provide further evidence that when there is not sufficient access to data, human decisions suffer.

6. Expert Review

We also conducted an expert review with six experts who regularly conduct record linkage and work with PII (5-10

years of experience). Experts were volunteers recruited from a professional network of people conducting record linkage studies. All experts completed an abbreviated version of the *high limit* condition used for the controlled experiment. The experts then answered questions about the potential utility and limitations of the approach and system.

In their own work, five of the experts normally conducted record linkage with full access to PII. They perceived that this system offered more privacy protection, with little to no impact on accuracy in the linkage, but may take more time. One expert had prior experience using encryption-based methods of data hiding for private record linkage with no access to PII. This participant perceived our system to have less protection and require more time compared to the encryption-based method, but to also allow for much better accuracy. He stated "I never know how well the hashing worked, or how accurate it is. It would be helpful to use this method to spot check a random sample (e.g., 5%)". This seems to agree with our goal of providing a level of access between the all or nothing that provides better accuracy than no access, but more protection than full access.

Five experts felt that the on-demand method did impact their decision making, while one did not because "I felt like I didn't need to click on most of them because my comfort level wouldn't increase". He did not think seeing more information would alleviate the uncertainty in the decision anyway. This points to the fundamental difficulty of uncertainty working with real data and affirms that the meta-data presented via the visual masks had sufficient information to support good decisions. One noted, "It works well, but it is time consuming to make the decision on whether to open the information you need".

When asked about potential benefits for this method, four mentioned privacy protection, one mentioned better accuracy of linkages, and one mentioned less fatigue of the data worker. More specifically, one expert mentioned the increased protection from the ability to accurately measure how much data was accessed (transparency) during linkage while another expert mentioned that the ability for the data custodians to limit the amount of access (budget) as being a privacy benefit. The respondent who discussed less fatigue also stated that, "Once I got used to the coding, allowing partial disclosure helped in decision making", pointing to our goal of actually improving linkage (i.e., more consistent linkage decisions) by providing better processed information for decision making in place of raw data.

When asked about specific contexts in which this system is especially useful, four stated it is useful for linking sensitive high-risk data such as health data, where privacy protection was important (e.g., "especially when linking to patient-provided data and where unique identifiers are not available"). Overall, the feedback was promising for the future potential of this direction of work, though the comments about the cognitive load for thinking about what to open suggest the need for future research, good training, and more practice.

7. Discussion

Research has demonstrated that information privacy is a budget-constrained problem that requires reasoning about the *tradeoff* between privacy and utility for a given context [39-41, 49]. Consequently, there is no "one-size-fits-all" solution, and there is no way to benefit from using data without taking some privacy risks. Our research tackles this difficult problem of finding the "sweet spot" between accessing PII for legitimate use while providing the maximum privacy protection as possible through the privacy by design approach.

We designed a system that reduces privacy risk through on-demand incremental information disclosure, which facilitates data work while making partial details available "as needed". This on-demand disclosure facilitates a practical implementation of the legal "minimum necessary disclosure" and accountable access requirements that are core principles of the new GDPR and HIPAA regulations making it possible to find a realistic middle ground between access to *all or no* access to PII.

From the experiment of different types of feedback and access restrictions for on-demand disclosure, the results show that all three variations were effective in increasing protection by reducing unnecessary disclosure. First, on-demand interactive disclosure (*no meter* condition) was able to significantly reduce disclosure to only KAPR = 7.85% while still being able to maintain similar quality scores. This is significantly less than what was possible with only a static display (36.7% vs. 7.85%) [20]. The prevention of unnecessary PII from being disclosed during record linkage can prevent most of the incidental identifications by people they know (e.g., neighbors, friends, co-workers) that patients are concerned about. This is exactly the local privacy that was of the most concern to patients in a survey conducted at the Mayo Clinic. For many patients, "their greatest concern about privacy actually had to do with their privacy locally ...[A] neighbor ... may still sometime be able to see [my] protected health information in the course of their work" [42, 43]. Thus, the main threat model for this work is an insider threat. It is to protect patient confidentiality by preventing someone from accidentally learning about the health status of people they know when handling PII.

Second, given the near-significant results of p = 0.07 for KAPR scores with the feedback meter present, this motivates interest in further study of whether quantifying the risk ahead of disclosure to inform decisions to disclose certain PII may be effective. One potential reason that differences

were not large may be due to the fact that the on demand disclosure without the meter already had very low levels of disclosure at only 7.85%. Thus, there was not much room to go lower without impacting the decision quality. Regardless of the effect of the meter on reducing disclosure, it is important to remember that quantifying the actual disclosure and sharing it with the users has a more important role. As with surveillance cameras, recording, quantifying, and displaying the risk to users has the potential to keep insiders on good behavior. One limitation of our user study is that we needed to focus on the interface and were not able to study how effective the meter was on keeping people on good behavior because the scenario we used kept everyone on good behavior. Future studies are needed to understand how much logging of computer systems, audits, and reminders of these logs might discourage bad behavior.

In addition, by quantifying and recording exactly how much risk was involved in a particular study via the meter, we can now have transparency, accountability, and communications in the record linkage process. For example, if one linkage project was able to achieve good linkage at one level, but another required much higher levels of disclosure, compliance may investigate the reason. Furthermore, with agreed-upon quantification of risks, we can now have clear conversations about what level of disclosure is appropriate at a much granular level, as apposed to limiting the options to either "access to all PII" or "no access". This conversation may include iteratively increasing or decreasing disclosure as we learn along the way.

Finally, the impact of enforcing a pre-specified limit on the disclosure was more complex. Our study clearly supported the findings from a previous study [20] that when there is not sufficient information disclosure, the quality of the linkage decision suffers (*H3.2*). On the other hand, when a sufficiently large limit was provided, participants seemed to disclose a bit more compared to the condition with unlimited budget (7.87% vs. 5.33%), though most spent only a fraction of the budget provided (21.4%). The amount disclosed was not statistically different from the *unlimited meter* condition, though the study cannot support claims for equivalence. The quality score was also similar to the *unlimited meter* condition, which may indicate that the high limit budget may be near the minimum level of disclosure needed to achieve this level of accuracy scores in the given data. This might indicate that erring on setting higher limits might be more effective since participants may still choose not to disclose the most possible, especially when they know the disclosure is transparently recorded.

The main feedback from the experts was that the system facilitated safe linkage without compromising on the quality of the results proving a good balance between the all or nothing 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 statistical difference in completion time was found among the different modes. This is likely due to the fact that when we prevent users from looking at details that are not needed to increase privacy, there is a potential bonus benefit of streamlining the interface so that the users 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 have the benefit of better focused attention.

Interactive incremental disclosure that can support just-in-time decisions can be a powerful design mechanism to enhance privacy. We posit that it has the potential to have as wide an impact on privacy-enhanced systems as encryption, but inevitably the design has to be context dependent on the data task. More research is needed to understand exactly what data is needed for human decisions, when access decisions are best determined, and how to best partition access for different types of data tasks. Our findings clearly support the literature on designing better systems such as these to nudge better privacy behavior; designing systems from the beginning with privacy in mind and incorporating various interventions (e.g., education, feedback, incentives) into the system is the only way to enable safe use of sensitive data.

8. Conclusion

Research has demonstrated the detrimental effect of not allowing sufficient human access for data tasks [8, 11-13, 44, 45]. 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, in order to obtain high quality data and bias-free record linkage, human involvement is essential to fine tune the results from automated systems (e.g., parameter settings, setting cutoff thresholds, iterative data standardization, building training datasets, validating results) [6]. Human interaction means that some data, under some suitable conditions, must be revealed to trusted persons to produce accurate linkages.

Our research provides evidence that incremental disclosure can be highly effective for ensuring legal compliance with the "minimum necessary" and accountable access requirements. Further interdisciplinary research is needed to learn the best ways to integrate these different technologies into an optimal system for privacy and utility of personal data.

Acknowledgements

References

[1] Hye-Chung Kum, Ashok Krishnamurthy, Ashwin Machanavajjhala, and Stanley C Ahalt. Social genome: Putting big data to work for population informatics. *Computer*, 47(1):56–63, 2014.

[2] Robin E Clark, Mihail Samnaliev, Jeffrey D Baxter, and Gary Y Leung. The evidence doesn't justify steps by state medicaid programs to restrict opioid addiction treatment with buprenorphine. *Health Affairs*, 30(8):1425–1433, 2011.

[3] William H Fisher, Robin Clark, Jeffrey Baxter, Bruce Barton, Elizabeth O'Connell, and Gideon Aweh. Cooccurring risk factors for arrest among persons with opioid abuse and dependence: implications for developing interventions to limit criminal justice involvement. *Journal of substance abuse treatment*, 47(3):197–201, 2014.

[4] C Joy Stewart, Hye-Chung Kum, Richard P Barth, and Dean F Duncan. Former foster youth: Employment outcomes up to age 30. *Children and Youth Services Review*, 36:220–229, 2014.

[5] Hye-Chung Kum, Ashok Krishnamurthy, Ashwin Machanavajjhala, Michael K Reiter, and Stanley Ahalt. Privacy preserving interactive record linkage (PPIRL). *Journal of the American Medical Informatics Association*, 21(2):212–220, 2014.

[6] Hye-Chung Kum, Stanley Ahalt, and Darshana Pathak. Privacy-preserving data integration using decoupled data. In *Security and Privacy in Social Networks*, pp. 225-253. Springer, New York, NY, 2013.

[7] Hyunmo Kang, Lise Getoor, Ben Shneiderman, Mustafa Bilgic, and Louis Licamele. Interactive entity resolution in relational data: A visual analytic tool and its evaluation. *IEEE transactions on visualization and computer graphics*, 14(5):999–1014, 2008.

[8] Janet M Bronstein, Charles T Lomatsch, David Fletcher, Terri Wooten, Tsai Mei Lin, Richard Nugent, and Curtis L Lowery. Issues and biases in matching medicaid pregnancy episodes to vital records data: the arkansas experience. *Maternal and child health journal*, 13(2):250–259, 2009.

[9] Cathy J Bradley, Charles W Given, Zhehui Luo, Caralee Roberts, Glenn Copeland, and Beth A Virnig. Medicaid, medicare, and the michigan tumor registry: a linkage strategy. *Medical Decision Making*, 27(4):352–363, 2007.

[10] Francis P Boscoe, Deborah Schrag, Kun Chen, Patrick J Roohan, and Maria J Schymura. Building capacity to assess cancer care in the medicaid population in New York State. *Health services research*, 46(3):805–820, 2011.

[11] Ileana Baldi, Antonio Ponti, Roberto Zanetti, Giovannino Ciccone, Franco Merletti, and Dario Gregori. The impact of record-linkage bias in the cox model. *Journal of evaluation in clinical practice*, 16(1):92–96, 2010.

[12] Stacie B Dusetzina, Seth Tyree, Anne-Marie Meyer, Adrian Meyer, Laura Green, and William R Carpenter. Linking data for health services research: a framework and instructional guide. 2014.

[13] Partha Lahiri and Michael D Larsen. Regression analysis with linked data. *Journal of the American statistical association*, 100(469):222–230, 2005.

[14] Rob Hall and Stephen E Fienberg. Privacy-preserving record linkage. In *Privacy in statistical databases*, volume 6344, pages 269–283. Springer, 2010.

[15] Dinusha Vatsalan, Peter Christen, and Vassilios S Verykios. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969, 2013.

[16] Arvind Narayanan and Vitaly Shmatikov. *Myths and fallacies of personally identifiable information. Communications of the ACM*, 53(6):24–26, 2010.

[17] Giovanni Iachello, Ian Smith, Sunny Consolvo, Mike Chen, and Gregory D Abowd. Developing privacy guidelines for social location disclosure applications and services. In *Proceedings of the 2005 symposium on Usable privacy and security*, pages 65–76. ACM, 2005.

[18] Aritra Dasgupta and Robert Kosara. Adaptive privacypreserving visualization using parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2241–2248, 2011.

[19] Aritra Dasgupta, Min Chen, and Robert Kosara. Measuring privacy and utility in privacy-preserving visualization. In *Computer Graphics Forum*, volume 32, pages 35–47. Wiley Online Library, 2013.

[20] Eric D Ragan, Hye-Chung Kum, Gurudev Ilangovan, and Han Wang. Balancing privacy and information disclosure in interactive record linkage with visual masking. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 326. ACM, 2018.

[21] Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J Crandall, Roberto Hoyle, and Apu Kapadia. Viewer experience of obscuring scene elements in photos to enhance privacy. In *Proceedings of the 2018*

CHI Conference on Human Factors in Computing Systems, page 47. ACM, 2018.

[22] Serdar Çiftçi, Pavel Korshunov, Ahmet Oguz Akyuz, and Touradj Ebrahimi. Using false colors to protect visual privacy of sensitive content. In *Human Vision and Electronic Imaging Xx*, volume 9394, page 93941L. Spie-Int Soc Optical Engineering, 2015.

[23] Daphne Chang, Erin L Krupka, Eytan Adar, and Alessandro Acquisti. Engineering information disclosure: Norm shaping designs. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 587–597. ACM, 2016.

[24] Leslie K John, Alessandro Acquisti, and George Loewenstein. Strangers on a plane: Context-dependent willingness to divulge sensitive information. *Journal of consumer research*, 37(5):858–873, 2010.

[25] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. Nudges for privacy and security: understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, 50(3):44, 2017.

[26] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, et al. How does your password measure up? The effect of strength meters on password creation. In *USENIX Security Symposium*, pages 65–80, 2012.

[27] Ios Kotsogiannis, Ashwin Machanavajjhala, Michael Hay, and Gerome Miklau: Pythia: Data Dependent Differentially Private Algorithm Selection. In *Proceedings of International Conference on Management of Data*, pp. 1323-1337. ACM, 2017.

[28] Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering* 13, no. 6 (2001): 1010-1027.

[29] Latanya Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, *Fuzziness and Knowledge-Based Systems* 10, no. 05 (2002): 557-570.

[30] Josep Domingo-Ferrer, and Torra Vicenç. A critique of k-anonymity and some of its enhancements. In 2008 *Third International Conference on Availability, Reliability and Security*, pp. 990-993. IEEE, 2008.

[31] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. technical report. *SRI International*, 1998.

[32] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *22nd International Conference on Data Engineering (ICDE'06)*, pp. 24-24. IEEE, 2006

[33] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. *23rd International Conference on Data Engineering*, pp. 106-115. IEEE, 2007.

[34] Ninghui Li, Wahbeh H. Qardaji, and Dong Su. Provably private data anonymization: Or, k-anonymity meets differential privacy. *CoRR*, abs/1101.2604 49:55, 2011.

[35] Cynthia Dwork. Differential privacy. *Encyclopedia of Cryptography and Security*, pages 338–340, 2011.

[36] Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[37] Daniel J Weitzner, Harold Abelson, Tim Berners-Lee, Joan Feigenbaum, James Hendler, and Gerald Jay Sussman. Information accountability. *Communications of the ACM*, 51(6):82–87, 2008.

[38] Qinbo Li, Adam D'Souza, Cason Schmit, and Hye-Chung Kum. Increasing Transparent and Accountable Use of Data by Quantifying the Actual Privacy Risk in Interactive Record Linkage. Poster presentation at *Proceedings of the AMIA Symposium 2019*, Full technical report available on [arXiv:1906.03345 cs.DB] http://arxiv.org/abs/1906.03345

[39] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twentysecond ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210. ACM, 2003.

[40] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204. ACM, 2011.

[41] Tiancheng Li and Ninghui Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–526. ACM, 2009.

[42] Federal Trade Commission et al. *Innovations in health care delivery*, 2008.

[43] Daniel J Gilman and James C Cooper. There is a time to keep silent and a time to speak, the hard part is knowing which is which: Striking the balance between

privacy protection and the flow of health care information. 2009.

[44] Julia Lane and Claudia Schur. Balancing access to health data and privacy: a review of the issues and approaches for the future. *Health services research*, 45(5p2):1456–1467, 2010.

[45] Julia Lane. Optimizing the use of micro-data: An overview of the issues. *Trans. Data Privacy*, 23(3):299–317, 2007.

[46] Hanna Köpcke, Andreas Thor, and Erhard Rahm. Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, 3(1-2):484–493, 2010.

[47] Michael Stonebraker, Daniel Bruckner, Ihab F Ilyas, George Beskales, Mitch Cherniack, Stanley B Zdonik, Alexander Pagan, and Shan Xu. Data curation at scale: The data tamer system. In *CIDR*, 2013.

[48] Hye-Chug Kum, Eric Ragan, Gurudev Ilangovan, Mahin Ramezani. Soups data and analysis (https://github.com/pinformatics/soups2019) June 7, 2019.

[49] Hye-Chung Kum and Stanley Ahalt. Privacy-by-design: Understanding data access models for secondary data. *AMIA Summits on Translational Science Proceedings*, 2013:126, 2013.

[50] In re Estate of Broderick, 34 Kan. App. 2d 695, 703, 125 P.3d 564, 570 (2005)

[51] Jennifer Guthrie. 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. Annals Health L.. 2003;12:143.

[52] Schmidt v. U.S. Dep't 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)

[53] Article 5(1c) EU General Data Protection Regulation (GDPR)

[54] Cason Schmit, Kathleen Kelly, and Jennifer Bernstein. Cross Sector Data Sharing: Necessity, Challenge, and Hope, *Journal of Law, Medicine, & Ethics,* 47 S2 (2019). In press.

[55] https://www.hhs.gov/hipaa/for-professionals/faq/213/what-conditions-may-health-care-provider-use-entire-medical-record/index.html

[56] Willem G van Panhuis, Proma Paul, Claudia Emerson, John Grefenstette, Richard Wilder, Abraham J Herbst, David Heymann and Donald S Burke. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014;14:1144. Published 2014 Nov 5. doi:10.1186/1471-2458-14-1144

[57] Jerry L. Hintze and Ray D. Nelson. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, *52*(2), 181-184