International Journal of Medical Informatics 204 (2025) 106045

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Privacy-by-design: Case studies in interactive record linkage using a hybrid human-computer system

Hye-Chung Kum [a,b,*], Eric Ragan [c], Mahin Ramezani [a,b], Gurudev Ilangovan [a,b], Theodoros Giannouchos [a,d], Qinbo Li [a,b,e], Adam D'Souza [f,g], Elmer V. Bernstam [h], Jeffrey R. Curtis [i], Alva O. Ferdinand [a], Cason Schmit [a]

[a] *Population Informatics Lab, Department of Health Policy and Management, School of Public Health, Texas A&M University, College Station, TX, USA*
[b] *Department of Computer Science and Engineering; Texas A&M University, College Station, TX, USA*
[c] *Department of Computer & Information Science & Engineering, University of Florida, FL, USA*
[d] *Department of Health Policy & Organization, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, USA*
[e] *Meta, Seattle, WA, USA*
[f] *Centre for Health Informatics, University of Calgary, Calgary, AB, Canada*
[g] *Provincial Research Data Services, Alberta Health Services, Alberta, Canada*
[h] *School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA*
[i] *Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, 1825 University Blvd, Birmingham, AL, USA*

ARTICLE INFO

Keywords:
Record linkage
Interactive record linkage
Patient matching
Privacy-by-design
Data segmentation
Real world data

ABSTRACT

Objective: High-quality patient matching from several sources without a common identifier (ID) requires interactive record linkage (RL) using a hybrid human-computer system. MiNDFIRL (MInimum Necessary Disclosure For Interactive Record Linkage) is a hybrid prototype software system that facilitates maximizing linkage accuracy while minimizing information disclosure. We present and evaluate MiNDFIRL using two real-world case studies.

Materials and Methods: Two user studies were conducted linking 10,000 data pairs from EHR data and 18,240 unique patient IDs from patient generated data. After automated RL, manual review was conducted by three teams of four reviewers (12 total) using MiNDFIRL to resolve potential matches that required human judgment. Reviews for matches were conducted independently and disagreements were resolved through consensus. The teams then participated in a group discussion about MiNDFIRL using a semi-structured interview format.

Results and Discussion: The best algorithm, Random Forest, found 388 and 539 matches each for EHR and patient generated data algorithmically, but also output an additional 303 and 187 potential pairs that required manual review. 232 and 84 more matches were confirmed manually from these uncertain pairs respectively. Among the full uncertain pairs, only 30% of available identifying information was needed in MiNDFIRL to separate out 77% (232/303) and 45% (84/187) true linkages respectively. When available, first names and emails were the most frequently used fields in making RL decisions.

Conclusion: On-demand access and masking techniques along with risk quantification through a hybrid human-computer system can significantly reduce disclosure while still minimizing false positives and false negatives in real-world RL.

1. Introduction

Record linkage (RL) refers to the process of identifying records pertaining to the same individual across two or more databases without a common unique identifier [1–3]. Accurate RL is critical for aggregating data to fully realize the benefits of advanced analytics, including machine learning (ML) and artificial intelligence (AI). However, three key challenges persist: (1) the absence of common, error-free, unique iden-

* Corresponding author at: Population Informatics Lab, Department of Health Policy and Management; School of Public Health, Texas A&M University; College Station, TX 77843, USA.
E-mail address: kum@tamu.edu (H.-C. Kum).

https://doi.org/10.1016/j.ijmedinf.2025.106045
Received 3 December 2024; Received in revised form 6 July 2025; Accepted 8 July 2025

Available online 31 July 2025

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

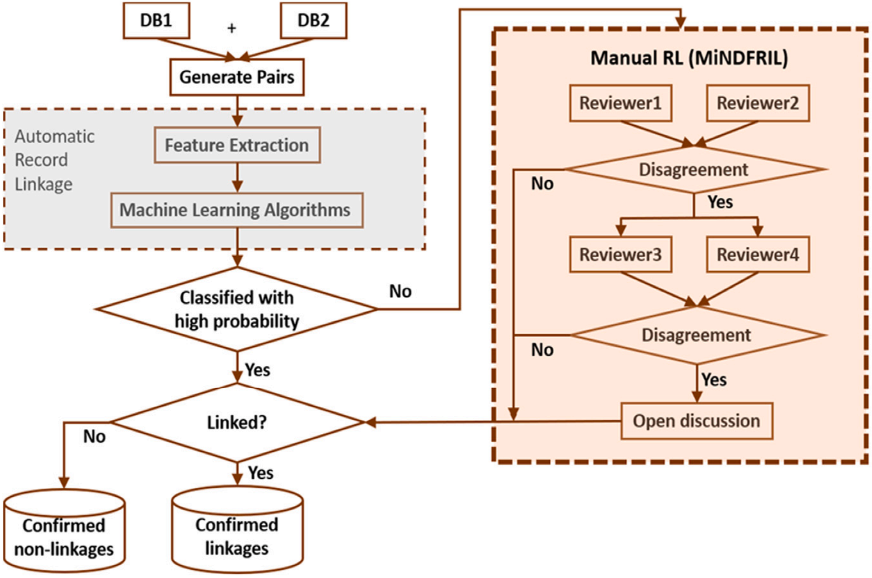


Fig. 1. Full Approximate RL Process Using a Hybrid Human-Computer System.

tifiers (e.g., medical record numbers, names) across data sources; (2) the necessity of human involvement to ensure high-quality RL; and (3) the requirement to use patient-level identifying information to achieve proper linkage. These challenges limit both the quantity and quality of studies using linked data, as well as their replicability.

Most prior research has focused either on improving automated RL methods to better handle messy data [4–6] or on developing privacy-preserving RL (PPRL) techniques to protect privacy [7]. A major gap remains: neither purely automated methods nor PPRL approaches adequately address the need for a *hybrid, interactive RL system* capable of meeting the quality standards required for scientific replicability. As a result, RL operations are often considered more of an art than a science, and few studies describe complete, end-to-end RL systems. This paper addresses that gap.

1.1. Privacy-by-design: interactive record linkage using a hybrid human-computer system

Substantial progress has been made in developing automated RL methods, including ML, deterministic, rule-based, and probabilistic approaches, which aim to account for the many variations in real data (e.g., variation in spelling, data errors, etc.) by allowing for approximate match, aka approximate RL [4–6]. The primary goal of approximate RL methods is to maximize true matches while minimizing false matches [8,9].

Many interactive RL systems implement a two-threshold approach: record pairs with similarity scores above the upper threshold are accepted as matches, those below the lower threshold are rejected as non-matches, and pairs with scores between the two thresholds are flagged for human review [10,11]. Uncertain pairs may either remain unlinked—thus reducing recall—or be manually adjudicated [4, 5]. Manual review is often considered essential in contexts where false-positive linkages (i.e., linking records from different individuals) could have serious consequences [12–14] or where fragmented data would impede critical tasks, such as post-marketing surveillance of adverse drug events.

Significant attention has also been given to privacy-preserving RL methods, such as hashing and secure multiparty computation [7]. However, these techniques often prevent evaluation of linkage quality or bounding of uncertainty. Balancing privacy protection with accurate RL in dynamic real-world settings requires a hybrid human–computer system capable of safely managing uncertainty. Consequently, many linkage centers iteratively refine RL algorithms and incorporate human reviewers to verify and correct potential linkages, a process known as *interactive RL*, which typically involves access to personally identifiable information (PII) [15,16].

Effective RL involving this type of manual review that requires PII often necessitates that multiple people access PII, which increases privacy risks including risk of identity theft or data leaks. Fig. 1 illustrates a full hybrid human–computer approximate RL system. This paper focuses on the manual review process (highlighted in the brown box) and aims to balance two goals by applying a *privacy-by-design* approach [8,17]:

- **Privacy goal:** Limit disclosure of identifying information (e.g., names) and guarantee no disclosure of sensitive information (e.g., diagnosis). Many legal protections, including the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule, apply specifically to individually identifiable health information [18,19].
- **Utility goal:** Sustain human effectiveness in making valid RL decisions, enabling evaluation of linked data quality and bounding of linkage uncertainties. This approach allows human judgment to balance false positives (i.e., incorrect links) and false negatives (i.e., missed correct links, resulting in fragmented data) based on project-specific requirements.

To our knowledge, this work is among the first to develop a hybrid human–computer interactive RL system that explicitly addresses both privacy and utility goals.

1.2. Objectives

MiNDFIRL (MInimum Necessary Disclosure For Interactive Record Linkage) is a prototype software system that meets these goals. We present and evaluate MiNDFIRL using two real-world case studies where we conducted two RL projects. We sought to identify new concerns and recommendations for consideration when moving MiNDFIRL from prototype to full application development. Specifically, this study sought to evaluate whether findings from prior formative investigations conducted in controlled user study environments would generalize to more complex, realistic operational settings. Areas of particular interest included the appropriateness of the interface design, the effects of the

The figure presents the main visual interface for MiNDFIRL. The key design elements are (1) minimum disclosure via interactive just-in-time interface by hiding data values, when possible, (Fig. 2b) and masking (i.e., adding visual meta-data such as icons and color coding to highlight discrepancies in data pairs) to help decision making without seeing raw data (Fig. 2c), (2) accountability via quantifying privacy risk (i.e., the privacy risk score on top) that allows for limiting privacy risk via a budget (i.e., the solid red line in privacy meter bar on top). See Appendix A for details on the KAPR (k-Anonymity Privacy Risk) Score. Abbreviations: DoB, date of birth; FFreq, first name frequency; ID, identifier; LFreq, last name frequency [22].

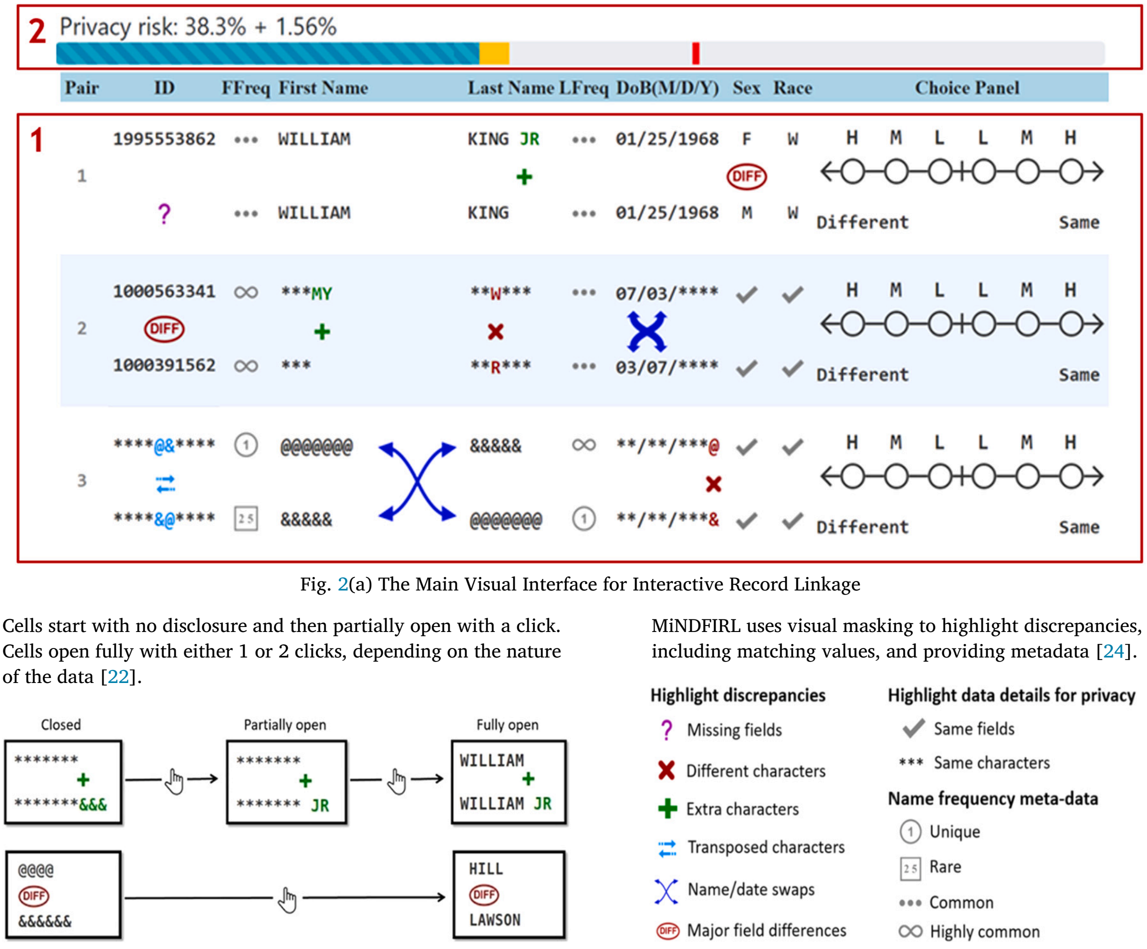


Fig. 2(a) The Main Visual Interface for Interactive Record Linkage

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 [22].

Fig. 2(b) Interactive On-Demand Interface

MiNDFIRL uses visual masking to highlight discrepancies, including matching values, and providing metadata [24].

Fig. 2(c) Visual Masking Icons

Fig. 2. MiNDFIRL prototype software [23]. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

on-demand disclosure technique on information access decisions, and the adaptability of MiNDFIRL to different data types (e.g., electronic health record [EHR] data versus patient-generated data collected via smartphone applications), all within the context of an end-to-end data cleaning and linkage pipeline. Whereas previous formative evaluations emphasized quantitative performance measures, the present summative evaluation prioritizes qualitative insights derived from participant experiences and feedback, providing a more holistic understanding of feasibility and implementation challenges. Data were collected through case studies involving small groups of data reviewers at two distinct sites, with the primary themes presented in the Discussion section.

2. MiNDFIRL (MInimum necessary disclosure for interactive record linkage)

MiNDFIRL is a hybrid human-computer prototype software system that effectively implements the "data minimization" ethical principle for privacy protection. The principle generally means that—in any data collection, disclosure, or use—data should be limited to the minimum necessary for the given purpose. Indeed, legal standards for disclosure in many data privacy laws (including the HIPAA Privacy Rule [20] and the General Data Protection Regulation (GDPR) [21]) are often limited to the "minimum necessary" to accomplish an intended purpose. These flexible standards permit full disclosure of available data when necessary (e.g., accurate RL). The full interface as seen by the user is shown in Fig. 2(a) which decomposes the interface to depict the two key design elements discussed next [22].

2.1. *Minimum disclosure via interactive just-in-time interface and masking*

In summary, the developed techniques manage privacy and data availability through software designed to limit the disclosure of personal details only on an "as-needed" basis based on data segmentation techniques [25,26] while supporting accountability by recording data-access events [22]. The method is complementary to automated RL prior to human involvement, and the software's graphical user interface (Fig. 2(a)) employs visual data masking to limit the amount of raw data available by default for human review [24]. Informative icons and visual highlighting (Fig. 2(c)) 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 (Fig. 2(b)) 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.

2.2. Accountability via quantified privacy risk and limiting the privacy budget

Risk quantification is important to support transparency, reasoning, communication, and decisions on the privacy and utility trade-off. Our system aims to quantify identity disclosure risk (sensitive attribute disclosure is negated by keeping the sensitive attributes separate from the identifying attributes). Thus, our prototype developed and used the k-Anonymity Privacy Risk (KAPR) score defined below, which uses the anonymity-set size as an estimate of the identity disclosure risk. *Anonymity-set size* measures privacy risk by counting how many individuals share identical identifiers; larger sets mean lower risk. Disclosing common names (e.g., John) poses lower identification risk, whereas a rare name (e.g., Tragedeigh) might pose much greater risk. Anonymity-set size is easily calculated dynamically with system interactions. For example, disclosing more information to aid linkage decreases the set size and raises privacy risks. The KAPR score ranges from 0% (no disclosure) to 100% (full disclosure), increasing with both the amount and uniqueness (defined by datasets being linked) of disclosed information. See Appendix A for a demonstrative example. Despite the KAPR score's effectiveness in our user study, the specific function's choice is less critical than employing a clear privacy risk feedback method for users.

Definition 2.1 *(k-Anonymized Privacy Risk (KAPR) Score)*. Let N be the full number of records across the databases being linked with D attributes that were used to build potential pairs for review. Let $\mathfrak{X}$ be the information disclosure state associated with a partial display $\mathcal{X}$ with $2n$ records (i.e., 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 *k-Anonymized Privacy Risk score* (KAPR) is given by

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

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

MiNDFIRL's configuration options allow for additional disclosure limits that could be applicable to the data—such as a law or data use agreement (DUA) that restricts access to only 4 digits of a Social Security number (SSN). This privacy budget, which is typically established a priori, may be useful to discuss with governing institutional review boards (IRB) at the start of a project, and in certain cases, might be set to a low enough threshold that preserves the description of a dataset subjected to this process as a Limited Data Set or even de-identified data. These MiNDFIRL configuration options permit project managers to tailor manual reviewers' privacy budgets for different data projects, including enforcing a limit on the total permissible disclosure for a given use case. The capability to pre-specify and enforce a privacy budget (i.e., disclosure limit) ahead of time provides assurances to data custodians and promotes trust among project partners. Fundamentally, the aim of legitimately accessing sensitive data is to optimize its usefulness under a fixed privacy budget. However, determining an appropriate privacy budget for a given task to support quality data is an open research area that will require further study.

2.3. Prior results

Much of the user interface for MiNDFIRL was designed based on two formative evaluations that provided empirical data for technique verification and established foundational knowledge for trade-offs among decision-making quality, privacy, and access behaviors using interactive on-demand techniques with a privacy budget to limit total access [22,24]. For example, results indicated how different privacy limits might lead to different human behavior in making decisions to click for more information, as well as how these limits on the privacy score impact the quality of the RL task. However, the formative studies were conducted with the software operating on small test data sets to support user interface design.

A comprehensive study of the ML based automated RL algorithms available in MiNDFIRL, the grey box in Fig. 1, are presented in Ramezani et al. [4]. Four models were trained and evaluated on EHR data: Random Forest (RF), Radial SVM, Linear SVM, and a Dense Neural Network (DNN). These models were trained using a diverse set of engineered features, including string similarity metrics (e.g., Jaro-Winkler, Damerau-Levenshtein, LCS), name embeddings (Name2Vec), birthdate components, and rule-based indicators such as name swaps and gender combinations. Among the tested models, the Random Forest model yielded the highest F1 score (0.992) with very few false positives (FP $=$ 1) and was selected for use in MiNDFIRL due to its strong balance between precision and recall. Full model specifications, tuning parameters, and evaluation results are publicly available on GitHub [27,28].

In addition, we conducted several studies [23,29,30] with patients and IRB members to develop template documents that best communicate the key components of the software to relevant stakeholders using tutorials [28]. These documents (i.e., privacy statement in the form of FAQ for the public, IRB application template, and DUA template) are released with the open-source software on GitHub and can be adapted as appropriate on projects [28]. In this paper, we integrate all these components to conduct two user studies using real world data in two different settings and report results of the summative evaluation of MiNDFIRL.

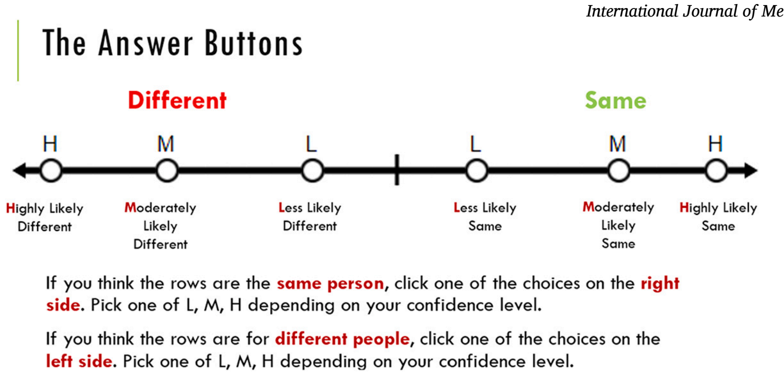
3. Materials and methods

3.1. Study design

The study included two separate case studies at two academic institutions, the University of Texas at Houston (UTH) and the University of Alabama at Birmingham (UAB), linking real data sets sampled from the corresponding locations. The studies consisted of linkage projects with teams involving four data reviewers with one of the reviewers also having a second role as team manager. The data reviewers used MiNDFIRL to review data discrepancies to perform RL. The team manager was responsible for configuring the software for allowable data fields, setting the privacy budget for data access, and assigning data pairs to the reviewers. The process of setting up the linkage projects and configuring MiNDFIRL served as proof of concept for applying the software techniques and integrating into real data environments with specific data needs and properties (e.g., which data fields, how many fields, which data to link, and coordination between human RL processes and automated RL methods). Where applicable, all settings used the recommended default values based on the user studies. The study protocol was approved by the IRB of respective institutions for each study.

The software included a tutorial explaining the interface and RL task. The data reviewers were tasked with reviewing the assigned sets of discrepant pairs. For each pair, participants were required to indicate whether the two patients should be considered the same or different individuals (i.e., did the two records belong to the same physical patient?), and each linkage decision included a level of confidence (low, medium, or high) (see Fig. 2a). To minimize different interpretations, the tutorial included guidance on how to respond as shown in Fig. 3.

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 two separate sessions over a period of one week. Session times varied, with linkage sessions taking approximately 60 to 90 minutes for the first session (Fig. 4, reviewer1 and reviewer2) and 30 to 40 minutes for the second session (Fig. 4, 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



The Answer Buttons

If you think the rows are the **same person**, click one of the choices on the **right side**. Pick one of L, M, H depending on your confidence level.

If you think the rows are for **different people**, click one of the choices on the **left side**. Pick one of L, M, H depending on your confidence level.

Fig. 3. Tutorial on guidance on how to respond.

questionnaires after each work period. 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.

Since the goal was to achieve the most accurate RL, the user study included an additional step to resolve differences in the data review. After all team members completed their linkage assignments, the software flagged pairs that were inconclusive among team members (i.e., any time two reviewers indicated two data rows corresponded to the same entity while the other two reviewers indicated they were different entities). Each team then participated in a “conflict-resolution” discussion (Fig. 4, “Open discussion”) in which the team members all reviewed the discrepant cases together. To facilitate this discussion, the software provides a special viewing mode that allows team members to review cases and see the responses of other team members. The manager led teams in discussion of discrepant cases together (synchronously, via video conferencing) until consensus was reached. 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 consensus could not be reached with limited data values, the manager could reference the complete records 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 semi-structured interview to collect feedback about (a) general system usage and processes, (b) strategies and decision-making with 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 understanding.

3.2. Participants

The user studies included a total of 12 data reviewers. The case study at UTH included 8 participants and integrated members of the research team with data reviewers. The study consisted of two teams of four (one manager and three data reviewers). The UAB case study included four participants working as a single team. Participant backgrounds and experience with data linkage varied but all of them had experience conducting record linkage on linkage projects. On one end it included experts as those developing record linkage algorithms while on the other end it included research assistants familiar with data management, linkage projects, and using linked data.

3.3. Data configuration

Both user 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) [5] on a Linux system. The benchmark data were developed from 10 million record pairs generated from the electronic health record (EHR) patient database by six reviewers who manually reviewed 20,000 randomly selected, potential match record-pairs to identify matches as detailed in [5]. For the linkage activity, eight fields were included: first name, middle name, last name, date of birth, SSN, gender, address, and phone number.

The second case study at UAB used data from rheumatology patients who had enrolled in the ArthritisPower patient powered research network (PPRN) registry (since renamed PatientSpot [33]). PatientSpot was previously one of the twenty PPRNs created by the national Patient-Centered Clinical Research Network (PCORnet). PatientSpot had collected 18,240 unique patient IDs on its smartphone app and web-based platform that were made available to the project on a secure Windows server. The fields used for the linkage study included registry 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 (NPI) (although not all fields were required).

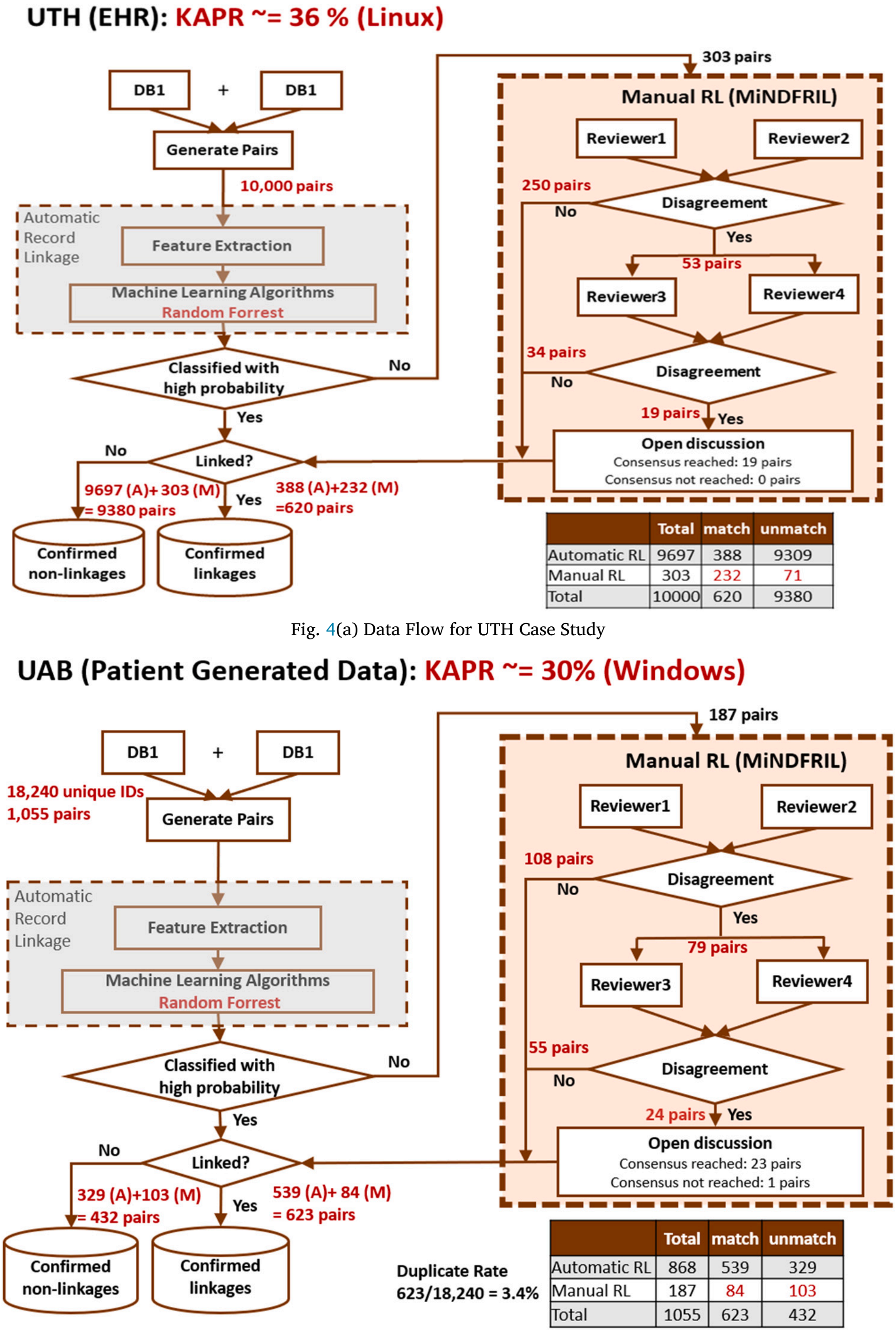
4. Results

Figs. 4(a) and (b) depict the full data flow for the 2 studies. Both studies were deduplication studies linking the same data to itself to identify duplicate records. Participants had no technical challenges getting MiNDFIRL to run on both Linux and Windows systems as well as using it on both EHR and patient-generated data. In the UTH study, starting with the 20,000 labeled pairs, our study team used 10,000 labeled pairs to conduct a comprehensive study of different ML models [4]. MiNDFIRL incorporates multiple ML algorithms, and overall, the random-forest (RF) linkage model performed best in terms of recall, F1-score, and minimum number of uncertain pairs. Thus, we used the RF model to study the full linkage process with the other 10,000 labeled pairs. The RF model found 388 matches, 9697 non-matches, and 303 potential matches that required human judgment. In the UAB study, 1055 unique pairs were generated 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 RF trained model from the UTH study for automatic RL on these pairs and adjusted the manual review thresholds. The model found 539 matches, 329 non-matches, and 187 potential matches that required human judgment.

Most of the manually reviewed pairs were adjudicated by two independent data reviewers with no disagreement in the first review. The pairs with disagreement (i.e., no match) were then reviewed by two

The following figure depicts the full flow for the hybrid MiNDFIRL System starting with generating the pairs from the two datasets at the top left being linked. This is followed by the grey box indicating the first step of conducting the automated RL process. Pairs that are output from the automated RL process with high probability of being linked or not being linked are confirmed appropriately, while those that are output as uncertain pairs are reviewed manually by 4 reviewers using the process depicted on the right. The results of both the automated RL (on the left), and the manual RL (on the right) are combined for final results depicted in the table on the bottom right.

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; UTH, University of Texas Health Science Center at Houston. UAB, University of Alabama at Birmingham Health System



	Total	match	unmatch
Automatic RL	9697	388	9309
Manual RL	303	232	71
Total	10000	620	9380

Fig. 4(a) Data Flow for UTH Case Study

	Total	match	unmatch
Automatic RL	868	539	329
Manual RL	187	84	103
Total	1055	623	432

Fig. 4(b) Data flow for UAB case study

Fig. 4. Data Flow for the Hybrid MiNDFIRL System.

more independent data reviewers (Fig. 4). Three of the four reviewers agreed on most of these pairs, leaving only 19 pairs (UTH) and 24 pairs (UAB), 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 team manager. In total, 232/620 (37%) and 84/623 (13%) more matches were found through the manual review process, but this required separating these 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 RF 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. Both UTH and UAB reviewers were able to achieve satisfactory linkage only disclosing the default KAPR setting in MiNDFIRL (i.e., 30% based on our findings from 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 most often disclosed by the review team during manual linkage, with most of the budget being spent on looking at email addresses.

5. Discussion

Our findings have several important implications for the RL practice and research. Critically, this study supports the use of on-demand disclosure techniques in interactive RL as an effective method to substantially limit PII viewed by human reviewers in real-world RL tasks. We also find that the design choices in the user interfaces for hybrid human-computer RL systems are crucial for efficient manual RL. Furthermore, we found that reviewers took different strategies for managing their personal “privacy budget” for revealing data details, which provides opportunities for future research and optimizing collaborations within RL teams. Lastly, our findings reinforce the notion that the concept of a definitive “gold standard” may be elusive in real-world applications, underscoring the need for flexible, context-aware evaluation frameworks. We discuss each of these implications in detail below.

5.1. On-demand disclosure is an effective approach to minimize privacy risks

On-demand disclosure of hidden data values in the context of interactive RL can limit the total amount of PII viewed by human reviewers. Prior published controlled experiments in [22,24] provided evidence that on-demand access and masking techniques and risk quantification are effective in significantly reducing the need to access identifying data. In sum, the MiNDFIRL interface reduced the k-Anonymized Privacy Risk (KAPR) Score to only 7.85% with little to no impact on RL accuracy or completion time [22]. KAPR Score uses the anonymity-set size as an estimate of the identity disclosure risk. The exact measure is given in Section 1.2 and more details with a demonstrative example are given in Appendix A. In [24] we conducted a large scale controlled study for different conditions using the KAPR Score meter, namely not displaying a KAPR Score meter during RL, having a meter with unlimited budget, restricting the disclosure to a high limit for KAPR Score, and low limit for KAPR Score. We found that the exact quantification method matters less than the fact that it is being quantified and people doing the manual review are given continuous feedback on the increased risk as they disclose more information. This is to be expected as people will pay more attention to whether disclosure is needed when they are reminded of the increased risk. To the best of our knowledge, there are no other methods that allow for partial disclosure in record linkage. Thus, in other methods this is significantly less than the 100% with all data disclosed in a fully identifiable dataset.

The two case studies presented here serve as a proof-of-concept that the same techniques are effective in more realistic RL settings used in combination with ML based automated RL methods. Both UTH and UAB cases were able to achieve accurate human linkage with the default disclosure budget for identifying information of KAPR = 30%. Participant feedback did not indicate notable problems or limitations that could not be addressed by using MiNDFIRL’s disclosure techniques, though data reviewers 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. The software streamlined the process to help reviewers access the appropriate individual record(s) from the full source, as needed.

5.2. The interface design is crucial for efficient manual RL

Comments from the case study revealed agreement with and reinforced the findings of the interface design from prior controlled experiments. Though the two case studies used different specific fields, both study sites used the same general masking and highlighting methods. The feedback indicated that the visual highlighting of discrepancies and addition of icons were effective for helping reviewers identify differences between patients 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 need for data disclosure. 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 (e.g., character insertions or deletions, character swaps, field value swaps, whole-value differences).

Importantly, not all interface icons were equally useful. Some reviewers 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 helped determine the uniqueness of the name. To reduce information complexity provided to data reviewers, the interface provided frequency level as four 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 that the optimal frequency levels will depend on the specific 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.

5.3. Understanding differences in user RL strategies is important for optimal RL

As expected, we found that different reviewers adopted different strategies and mindsets when conducting RL. For example, certain reviewers might give more attention to an *ID* field, while other reviewers might put more weight on a *date of birth* field for making linkage decisions. While not a problem, this finding reinforces the importance of software that supports collaborative decision-making and disagreement resolution to address between-people differences and perspectives throughout the linkage process. Future iterations of software that support our method might explore the integration of algorithmic techniques that can help log the history of data-access preferences by individual data reviewers (e.g., allow reviewer A to see that reviewer 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 as well as training beginners.

Reviewers 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 for 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 assumes that reviewers will only access more details when necessary for improved decisions. It is important to note that users should be encouraged to review more data details only when they believe it will add value and improve accuracy. 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.

5.4. Accuracy remains difficult to assess using real-world data

One limitation of the studies is that in real data there is no clear "ground truth", so we are unable to provide definitive results for the effects on RL accuracy. However, there was very strong agreement among the 4 reviewers in both studies indicating that although algorithms may find these pairs uncertain, most humans do not. Initial agreement between 2 initial reviewers was 83% and 58% each for the UTH and UAB studies respectively, with discordant pairs able to be resolved by majority (3 of 4 reviewers) for an additional 11% and 28% respectively, yielding overall agreement of 94% and 86% respectively. Final consensus via discussion was reached for all but 1 pair in the UAB study which required a final determination by the UAB manager. Although these agreement rates are encouraging, they may overestimate true reliability, as percentage agreement does not account for chance agreement. More robust measures such as Cohen's Kappa could not be calculated due to missing annotations on match versus non-match classifications.

6. Conclusion

Introducing the "human" in the "loop" of the data analysis and exploration process to augment the computational prowess of automated ML processes is key to obtaining usable high-quality results. There are important challenges in terms of how to facilitate these interactions when handling sensitive data including what should be disclosed and how for the most effective data decision making as well as best privacy protection [37]. Information privacy research has shown mathematically that any disclosure of data leads to some privacy loss [31,32]. Thus, the goal of any private data analysis should be to achieve the maximum utility under a fixed privacy budget [17]. There is still much to understand about the tradeoffs between utility and privacy, and few tools are available to facilitate the tradeoff decisions. MiNDFIRL is one tool available to facilitate these decisions for interactive RL using on-demand access techniques using data segmentation, masking, and risk quantification. Both case studies demonstrated that these techniques were effective in significantly reducing the needed access to identifying data for high-quality patient matching using a hybrid human-computer system in real-world settings for both EHR and patient-generated data.

CRediT authorship contribution statement

Hye-Chung Kum: Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Eric Ragan:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Mahin Ramezani:** Writing – review & editing, Writing – original draft, Project administration, Investigation, Formal analysis. **Gurudev Ilangovan:** Writing – review & editing, Investigation. **Theodoros Giannouchos:** Writing – review & editing, Project administration, Investigation. **Qinbo Li:** Writing – review & editing, Investigation, Formal analysis. **Adam D'Souza:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis. **Elmer V. Bernstam:** Writing – review & editing, Resources, Funding acquisition. **Jeffrey R. Curtis:** Writing – review & editing, Resources, Funding acquisition. **Alva O. Ferdinand:** Writing – review & editing, Funding acquisition. **Cason Schmit:** Writing – review & editing, Writing – original draft.

Funding

Research reported in this article was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1602-34486). Dr. Bernstam's effort was also supported in part by NIH/NCATS grant number UL1 TR003167 and Dr. Curtis' effort was supported in part by NIAMS P30AR072583. The statements presented in this article are solely the responsibility of the authors and do not necessarily represent the views of the funders.

Declaration of competing interest

The authors have no competing interests to declare.

Acknowledgements

We would like to thank Jonathon Lowe for his help in final edits to the paper. We would also like to thank the patient participants of PatientSpot [33], a comprehensive disease platform and a patient registry. PatientSpot is supported as a joint collaboration between the nonprofit Global Healthy Living Foundation (GHLF) and the University of Alabama at Birmingham (UAB) and various other dedicated collaborators.

Appendix A. k-Anonymized privacy risk (KAPR) score

In order to minimize the privacy risk associated with access to PII, the presented on-demand disclosure method requires a method for risk quantification. In this section, we define a quantitative measure, called the *k-Anonymized Privacy Risk (KAPR) score*, for the identity disclosure risk associated with partial information disclosure. We illustrate by example how KAPR is calculated, and we state and prove the desired properties of the measure.

A.1. Preliminaries

The attributes occurring in a data table can be categorized into three non-disjoint sets:

1. *identifiers* (e.g. name, social security number), which directly identify a person;
2. *quasi-identifiers* (e.g. gender, ZIP code, race), which can be used to link records to external identifiers; and
3. *sensitive attributes* (e.g. disease status, income), whose values for any individual must not be disclosed.

A standard approach to preserving privacy is *deidentification*, which refers to removing the identifiers from the data before sharing. This is not sufficient to guarantee anonymity, as quasi-identifiers can potentially be used to link records to identify records in other data sets, thereby causing *identity disclosure* of individuals within the data set.

Table A.1
Example of display mode for different attributes.

Attribute	Type	Masked	Partial	Full
ID	String	*********@	*********9	1742682819
		*****&****	*****6****	1742668281
Name	Varchar	*******@@@	******* JR	SANCHEZ JR
		*******	*******	SANCHEZ
DOB	Date	@@/&&/****	08/09/****	08/09/1964
		&&/@@/****	09/08/****	09/08/1964
Race	Category	@	N/A	White
		&	N/A	Asian

We use a privacy risk score based on *k-anonymity* [34,35] to measure the risk of identity disclosure. This property places a restriction on the disclosure of identifiers and quasi-identifiers.

Definition A.1. A data set is said to satisfy *k-anonymity* for $k > 1$ if, for each combination of identifiers and quasi-identifiers, at least k records exist in the data set sharing that combination.

No person in a k-anonymous data set can be distinguished from at least $k-1$ other people, which helps guard against identity disclosure (in other words, no particular person's data can be identified with certainty as belonging to them). k-anonymity alone is not sufficient to protect against attribute disclosure (e.g., cancer status) [36]. However, when protection against attribute disclosure is guaranteed by separation of the sensitive attributes from the identifiers, it can be leveraged to create a useful measure for assessing the risk of identity disclosure (some amount of which is inherent in the record linkage context).

A.2. KAPR privacy risk score

Before defining the KAPR score, we formalize the concept of a *partial display*.

Definition A.2 *(partial display).* Let X be a data set with N records and D non-sensitive attributes, containing potential pairs of records to be linked. A partial display $\mathcal{X}$ is a set of $2n$ rows and D attributes obtained from $n \le \binom{N}{2}$ pairs of records from X that are displayed to a user performing interactive record linkage. The rows in $\mathcal{X}$ are many-to-one related to the records in X. The values of the attributes in $\mathcal{X}$ may be fully disclosed, partially disclosed or masked versions of the corresponding attributes in X, as described in Table A.1.

Underlying the partial displays is the closely related concept of an *information disclosure state*, which in turn requires the concepts of *proportion of characters disclosed* and *anonymity set size*.

Definition A.3 *(proportion of characters disclosed).* Let $\mathcal{X}$ be a partial display comprising $2n$ rows and D attributes. Let the number of characters of attribute j of record i be n_{ij}. Let the number of characters whose true values are disclosed be d_{ij}. Then, the *proportion of characters disclosed* for attribute j of record i is $p_{ij} := d_{ij}/n_{ij}$.

Definition A.4 *(anonymity set size).* The *anonymity set size* k_i for a row r_i in a partial display $\mathcal{X}$, based on an underlying data set X, is the number of records in X that could correspond to r_i, based on the level of disclosure in r_i.

Definition A.5 *(information disclosure state).* Let $\mathcal{X}$ be a partial display comprising $2n$ rows and D attributes. Let p_{ij} represent the proportion of characters of attribute j disclosed for row i. Let k_i be the size of the anonymity set of record i based on the information that has been disclosed. The *information disclosure state* of $\mathcal{X}$ is defined by the matrix $\mathfrak{X} \in [0,1]^{2n\times D}$, with matrix elements $\mathfrak{X}_{ij} := (p_{ij}/k_i) \in [0,1]$.

There is a many-to-one mapping from partial displays to information disclosure states. We can now define KAPR as a function of an information disclosure state.

Definition A.6 *(k-Anonymized Privacy Risk (KAPR) Score).* Let $\mathfrak{X}$ be the information disclosure state associated with a partial display $\mathcal{X}$ with $2n$ records and D attributes. 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 *k-Anonymized Privacy Risk score* (KAPR) is given by

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

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

A.3. Example

We illustrate the utility of KAPR via an example. Consider Table A.2, containing microdata about individual people, whose schema comprises the identifier {`Name`}, the quasi-identifiers {`DOB`, `Race`}, and the sensitive attribute {`Income`}. The data linkage task to be accomplished is to merge duplicate records in this data set. This is accomplished by having a person manually inspect all pairs of similar records and decide whether or not they correspond to the same individual. In this example, the ground truth is that records 1 and 2 belong to unique individuals, while records 3 and 4 correspond to the same individual (as indicated by the row colors). Because all of the records in this table are quite similar, the user interface will display every possible pair of records from this table to the reviewer.

Initially, the pairs will be displayed in masked mode. This corresponds to the information disclosure state depicted in Table A.3. The field `ID` contains a tuple denoting which records from the base data set Table A.3 constitute this pair, and the field i is used to label the rows of data displayed. Since `Income` is a sensitive attribute, it will not be displayed. We denote by $\mathbf{p}_i$ the vector whose entries consist of the proportions of characters of the `Name`, `DOB` and `Race` fields respectively. We denote by k_i the size of the anonymity set for record i. In principle, any number of the characters of `Name` and `DOB` can be disclosed. Because `Race` is a Category variable, it can only be not disclosed or fully disclosed; there is no option for partial disclosure. In this mode, each row i in the display has an anonymity set size of $k_i = 4$, because the masked data could correspond to any of the records in Table A.2. However, no characters of any of the three attributes are disclosed, meaning $\mathbf{p}_i = (0,0,0)$ for each row i, and hence $K_i = 0$ for each row as well. Thus, the KAPR score for this information disclosure state is 0.

The incremental display allows the user to reveal incremental additional information about any cell in the display, and the privacy cost of this action is quantified by the amount by which the KAPR score is incremented as result. As an example, the user may have decided to switch from masked mode to partial mode for every cell in the display, as depicted in Table A.4. This will change the values of k_i and $\mathbf{p}_i$ for each row. For example, for row $i = 1$, the size of the anonymity set k_i has decreased from 4 in masked mode to 3 in partial mode, since we now know that the last character of `Name` is ‘y’, which means it must correspond to one of the three rows where `Name` = ‘Mary’ in Table A.2. Furthermore, we have now revealed $1/4$ of the characters of `Name`, but still no characters of the other attributes, meaning $\mathbf{p}_i = (1/4, 0, 0)$ now. In this case, since $N = 12$, $D = 3$, this means that $K_i = 1/432$. The value of K_i for each row is shown separately; for $i \in \{11, 12\}$, $K_i = 0$ still, because the identifying attributes of the underlying records are identical, and thus no characters need be revealed. The KAPR score for this information disclosure state is $\sum_{i=1}^{12} K_i = 31/432 \approx 0.072 > 0$. So the KAPR score has increased incrementally from 0 when every row was displayed in masked mode. It should be clear from the example that the KAPR

Table A.2
Fully disclosed individual-level attributes. *Note:* Colors indicate individuals (ground truth). There are three different individuals; rows 3 and 4 correspond to the same individual.

ID	Name	DOB	Race	Income
1	Mary	08/09/1964	Hispanic	69,426
2	Mark	08/09/1964	Hispanic	38,001
3	Mary	09/08/1964	Black	27,998
4	Mary	09/08/1964	Black	27,989

Table A.3
Masked pairs.

ID	i	Name	DOB	Race	k_i	$\mathbf{p}_i$	K_i
(1,2)	1	***@	**/**/****	N/A	4	(0, 0, 0)	0
	2	***&	**/**/****	N/A	4	(0, 0, 0)	0
(1,3)	3	****	*@/*&/****	N/A	4	(0, 0, 0)	0
	4	****	*&/*@/****	N/A	4	(0, 0, 0)	0
(1,4)	5	****	*@/*&/****	N/A	4	(0, 0, 0)	0
	6	****	*&/*@/****	N/A	4	(0, 0, 0)	0
(2,3)	7	***@	*@/*&/****	N/A	4	(0, 0, 0)	0
	8	***&	*&/*@/****	N/A	4	(0, 0, 0)	0
(2,4)	9	***@	*@/*&/****	N/A	4	(0, 0, 0)	0
	10	***&	*&/*@/****	N/A	4	(0, 0, 0)	0
(3,4)	11	****	**/**/****	N/A	4	$(0,0,0)$	0
	12	****	**/**/****	N/A	4	$(0,0,0)$	0
						$K\ (=\sum_i K_i)$	0

Table A.4
Partial disclosure.

ID	i	Name	DOB	Race	k_i	$\mathbf{p}_i$	K_i
(1,2)	1	***y	**/**/****	N/A	3	(1/4, 0, 0)	1/432
	2	***k	**/**/****	N/A	1	(1/4, 0, 0)	1/144
(1,3)	3	****	*8/*9/****	N/A	1	(0, 2/8, 0)	1/144
	4	****	*9/*8/****	N/A	2	(0, 2/8, 0)	1/288
(1,4)	5	****	*8/*9/****	N/A	1	(0, 2/8, 0)	1/144
	6	****	*9/*8/****	N/A	2	(0, 2/8, 0)	1/288
(2,3)	7	***k	*8/*9/****	N/A	1	(1/4, 2/8, 0)	1/72
	8	***y	*9/*8/****	N/A	2	(1/4, 2/8, 0)	1/144
(2,4)	9	***k	*8/*9/****	N/A	1	(1/4, 2/8, 0)	1/72
	10	***y	*9/*8/****	N/A	2	(1/4, 2/8, 0)	1/144
(3,4)	11	****	**/**/****	N/A	3	$(0,0,0)$	0
	12	****	**/**/****	N/A	3	$(0,0,0)$	0
						$K\ (=\sum_i K_i)$	0.072

Table A.5
Full disclosure.

ID	i	Name	DOB	Race	k_i	$\mathbf{p}_i$	K_i
(1,2)	1	Mary	08/09/1964	Hispanic	1	$(1,1,1)$	1/12
	2	Mark	08/09/1964	Hispanic	1	$(1,1,1)$	1/12
(1,3)	3	Mary	08/09/1964	Hispanic	1	$(1,1,1)$	1/12
	4	Mary	09/08/1964	Black	2	$(1,1,1)$	1/24
(1,4)	5	Mary	08/09/1964	Hispanic	1	$(1,1,1)$	1/12
	6	Mary	09/08/1964	Black	2	$(1,1,1)$	1/24
(2,3)	7	Mark	08/09/1964	Hispanic	1	$(1,1,1)$	1/12
	8	Mary	09/08/1964	Black	2	$(1,1,1)$	1/24
(2,4)	9	Mark	08/09/1964	Hispanic	1	$(1,1,1)$	1/12
	10	Mary	09/08/1964	Black	2	$(1,1,1)$	1/24
(3,4)	11	Mary	09/08/1964	Black	2	$(1,1,1)$	1/24
	12	Mary	09/08/1964	Black	2	$(1,1,1)$	1/24
						$K\ (=\sum_i K_i)$	0.750

score for this information disclosure state does not depend upon the order in which the information was disclosed; regardless of the order in which the user clicked the cells of the display to reveal more information, the final information disclosure state is the same, and the KAPR score depends only on the information disclosure state.

The maximum amount of information that can be revealed is for every row to be displayed in full disclosure mode. This information disclosure state is depicted in Table A.5. In this state, $\mathbf{p}_i = (1,1,1)$ for every row i. Even in this fully disclosed state, $k_i = 2$ for some rows (the ones where the underlying record is one of the two duplicates in the original record set with respect to the identifying attributes). As a result, the KAPR score is 0.750, which is less than the theoretical maximum value of 1 for a data set without duplicates.

A.4. Properties of KAPR score

KAPR has several appealing and easily demonstrated properties that make it a useful measure of identity disclosure risk for interactive record linkage.

Property A.1. *The KAPR score $K : \mathbb{R}^{N\times D} \to \mathbb{R}_+$ is a norm (it is non-negative, homogeneous, and obeys the triangle inequality). Hence, it can be assigned the geometrical interpretation of the "length" of the information disclosure state, where "longer" means "higher risk of identity disclosure".*

Proof. This follows directly from Definition A.6 and the fact that $\kappa/ND > 0$, since a norm multiplied by a positive constant is still a norm (it represents a uniform rescaling of all lengths). □

Property A.2. *KAPR score manifestly depends only on the information disclosure state, which in turn depends only on the information displayed to the user, and not on the precise sequence of actions taken by the user to lead to that state.*

Property A.3. *KAPR score ranges from 0 (when no characters are disclosed, such as when every attribute is in masked display mode) to 1 (when every character is disclosed and the anonymity set size of every record in the underlying data set is 1).*

Proof. $K \geq 0$ follows from KAPR being a norm. $K \leq 1$ follows from the fact that $k_i \geq \kappa, |p_{ij}| \leq 1 \Rightarrow |p_{ij}|/k_i \leq \sum_{i=0}^{N-1} \sum_{j=0}^{D-1} 1/\kappa = ND/\kappa \Rightarrow K(\kappa; X) = (\kappa/ND)\|X\|_{1,1} \leq 1$. □

Property A.4. *KAPR score monotonically increases as more characters are disclosed.*

Proof. Consider the partial derivative of $K(\kappa, X)$ with respect to any of the p_{ij}: $\frac{\partial K}{\partial p_{ij}} = \frac{\kappa}{ND} \sum_{r,s} \frac{1}{k_r} \frac{\partial |p_{ij}|}{\partial p_{rs}} = \frac{\kappa}{ND} \sum_{r,s} \frac{1}{k_r} \delta_{ir} \delta_{js} = \frac{\kappa}{ND} \frac{1}{k_i} > 0$. Thus, KAPR always increases with disclosure of new characters. □

Property A.5. *The KAPR score penalty for disclosing attribute values for any particular record is a monotonically decreasing function of the anonymity set size for the record.*

Proof. Suppose characters from attribute j of record i are disclosed. From the proof of Property A.4 above, $\frac{\partial K}{\partial p_{ij}} = \frac{\kappa}{ND} \frac{1}{k_i}$. This is a decreasing function of k_i, the anonymity set size of record i. Moreover, disclosure can only decrease or maintain the anonymity set size, so a given disclosure that causes a reduction in the anonymity set size causes a greater increase in KAPR than the same disclosure would if the anonymity set size was preserved. □

Data availability

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

References

[1] S.B. Dusetzina, S. Tyree, A-M. Meyer, A. Meyer, L. Green, W.R. Carpenter, Background and Purpose. Linking Data for Health Services Research: A Framework and Instructional Guide [Internet], 2014.

[2] A.K. Elmagarmid, P.G. Ipeirotis, V.S. Verykios, Duplicate record detection: a survey, IEEE Trans. Knowl. Data Eng. 19 (1) (2006) 1–16.

[3] S.L. DuVall, A.M. Fraser, K. Rowe, A. Thomas, G.P. Mineau, Evaluation of record linkage between a large healthcare provider and the Utah population database, J. Am. Med. Inform. Assoc. 19 (e1) (2012) e54–e59.

[4] M. Ramezani, G. Ilangovan, H.C. Kum, Evaluation of machine learning algorithms in a human-computer hybrid record linkage system, in: A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021), Stanford University, Palo Alto, California, USA, March 22-24, 2021, 2021, CEUR Workshop Proc. 2846 (4) (2021).

[5] E. Joffe, M.J. Byrne, P. Reeder, et al., A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation, J. Am. Med. Inform. Assoc. 21 (1) (2014) 97–104.

[6] N. Barlaug, J.A. Gulla, Neural networks for entity matching: a survey, ACM Trans. Knowl. Discov. Data 15 (3) (2021) 1–37.

[7] A. Gkoulalas-Divanis, D. Vatsalan, D. Karapiperis, M. Kantarcioglu, Modern privacy-preserving record linkage techniques: an overview, IEEE Trans. Inf. Forensics Secur. 16 (2021) 4966–4987.

[8] H-C. Kum, A. Krishnamurthy, A. Machanavajjhala, M.K. Reiter, S. Ahalt, Privacy preserving interactive record linkage (PPIRL), J. Am. Med. Inform. Assoc. 21 (2) (2014) 212–220.

[9] S.C. Weber, H. Lowe, A. Das, T. Ferris, A simple heuristic for blindfolded record linkage, J. Am. Med. Inform. Assoc. 19 (e1) (2012) e157–e161.

[10] I.P. Fellegi, A.B. Sunter, A theory for record linkage, J. Am. Stat. Assoc. 64 (328) (1969) 1183–1210.

[11] S.L. DuVall, R.A. Kerber, A. Thomas, Extending the Fellegi–Sunter probabilistic record linkage method for approximate field comparators, J. Biomed. Inform. 43 (1) (2010) 24–30.

[12] J.M. Bronstein, C.T. Lomatsch, D. Fletcher, et al., Issues and biases in matching medicaid pregnancy episodes to vital records data: the Arkansas experience, Matern. Child Health J. 13 (2009) 250–259.

[13] A.B. McCoy, A. Wright, M.G. Kahn, J.S. Shapiro, E.V. Bernstam, D.F. Sittig, Matching identifiers in electronic health records: implications for duplicate records and patient safety, BMJ Qual. Saf. 22 (3) (2013) 219–224.

[14] I. Baldi, A. Ponti, R. Zanetti, G. Ciccone, F. Merletti, D. Gregori, The impact of record-linkage bias in the Cox model, J. Eval. Clin. Pract. 16 (1) (2010) 92–96.

[15] H. Kang, L. Getoor, B. Shneiderman, M. Bilgic, L. Licamele, Interactive entity resolution in relational data: a visual analytic tool and its evaluation, IEEE Trans. Vis. Comput. Graph. 14 (5) (2008) 999–1014.

[16] H. Köpcke, A. Thor, E. Rahm, Evaluation of entity resolution approaches on real-world match problems, Proc. VLDB Endow. 3 (1–2) (2010) 484–493.

[17] A. Narayanan, V. Shmatikov, Myths and fallacies of “personally identifiable information”, Commun. ACM 53 (6) (2010) 24–26.

[18] C. Schmit, B.N. Larson, H-C. Kum, Data privacy in the time of plague, Yale J. Health Pol'y L. & Ethics 21 (2022) 152.

[19] 45 CFR § 160.103, 2013.

[20] 45 CFR § 164.502(b), 2013.

[21] GDPR, Art. 5(1)(c), 2018.

[22] H.C. Kum, E.D. Ragan, G. Ilangovan, et al., Enhancing privacy through an interactive on-demand incremental information disclosure interface: applying {privacy-by-design} to record linkage, in: Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019), 2019, pp. 175–189.

[23] C. Schmit, A.O. Ferdinand, T. Giannouchos, H.-C. Kum, Case study on communicating with research ethics committees about minimizing risk through software: an application for record linkage in secondary data analysis, JAMIA Open (ISSN 2574-2531) 7 (1) (2024) ooae010.

[24] E.D. Ragan, H.C. Kum, G. Ilangovan, et al., 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, 2018, pp. 1–12.

[25] A. Grando, D. Sottara, R. Singh, A. Murcko, H. Soni, T. Tang, N. Idouraine, M. Todd, M. Mote, D. Chern, C. Dye, Pilot evaluation of sensitive data segmentation technology for privacy, Int. J. Med. Inform. 138 (2020) 104121.

[26] M.M. Goldstein, A.L. Rein, M.M. Heesters, P.P. Hughes, B. Williams, S.A. Weinstein, Data segmentation in electronic health information exchange: Policy considerations and analysis, U.S. Department of Health and Human Services, Office of the National Coordinator for Health IT, Washington, D.C., 2010, Report.

[27] Population Informatics Lab, hybridRL, https://github.com/pinformatics/hybridRL_code_and_models, 2020. (Accessed 30 November 2023).

[28] Population Informatics Lab, MINDFIRL, https://github.com/pinformatics/mindfirl_2025, 2025. (Accessed 17 August 2025).

[29] C. Schmit, K.V. Ajayi, A.O. Ferdinand, et al., Communicating with patients about software for enhancing privacy in secondary database research involving record linkage: delphi study, J. Med. Internet Res. 22 (12) (2020) e20783.

[30] T.V. Giannouchos, A.O. Ferdinand, G. Ilangovan, 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. 28 (8) (2021) 1746–1755.

[31] I. Dinur, K. Nissim, Revealing information while preserving privacy, in: Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2003, pp. 202–210.

[32] H-C. Kum, A. Krishnamurthy, A. Machanavajjhala, S.C. Ahalt, Social genome: putting big data to work for population informatics, Computer 47 (1) (2013) 56–63.

[33] Patientspot Home, Secondary patientspot home, https://patientspot.org/, 2020. (Accessed 30 November 2023).

[34] P. Samarati, Protecting respondents identities in microdata release, IEEE Trans. Knowl. Data Eng. 13 (6) (2001) 1010–1027.

[35] L. Sweeney, k-anonymity: a model for protecting privacy, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10 (05) (2002) 557–570.

[36] J. Domingo-Ferrer, V. Torra, A critique of k-anonymity and some of its enhancements, in: 2008 Third International Conference on Availability, Reliability and Security, IEEE, 2008, pp. 990–993.

[37] H.-C. Kum, S. Bedrick, M.C. Weigle, Challenges in data science in the use of large-scale population datasets for scientific inquiry, in: Digital Ethology: Human Behavior in Geospatial Context, The MIT Press, ISBN 9780262378840, 2024, https://direct.mit.edu/book/chapter-pdf/2458093/c010400_9780262378840.pdf.