

Privacy preserving interactive record linkage (PPIRL)

Hye-Chung Kum,^{1,2,3,4} Ashok Krishnamurthy,^{1,3,5} Ashwin Machanavajjhala,⁶
Michael K Reiter,³ Stanley Ahalt^{1,3,5}

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/amiajnl-2013-002165>).

¹Population Informatics Research Group, Department of Computer Science, UNC-CH & Department of Health Policy and Management, Texas A&M Health Science Center, USA

²Department of Health Policy and Management, Texas A&M Health Science Center, College Station, Texas, USA

³Department of Computer Science, UNC-CH, Chapel Hill, North Carolina, USA

⁴Department of Pediatrics, College of Medicine Baylor Scott & White, Texas A&M Health Science Center, Temple, Texas, USA

⁵RENCI, UNC-CH, Chapel Hill, North Carolina, USA

⁶Department of Computer Science, Duke University, Durham, North Carolina, USA

Correspondence to

Dr Hye-Chung Kum,
Department of Health Policy
and Management, Texas A&M
Health Science Center, 1266
TAMU, College Station,
TX 77843, USA;
kum@srph.tamhsc.edu

Received 3 July 2013

Revised 5 September 2013

Accepted 6 October 2013

Published Online First

7 November 2013

ABSTRACT

Objective Record linkage to integrate uncoordinated databases is critical in biomedical research using *Big Data*. Balancing privacy protection against the need for high quality record linkage requires a human-machine hybrid system to safely manage uncertainty in the ever changing streams of chaotic *Big Data*.

Methods In the computer science literature, private record linkage is the most published area. It investigates how to apply a known linkage function safely when linking two tables. However, in practice, the linkage function is rarely known. Thus, there are many data linkage centers whose main role is to be the trusted third party to determine the linkage function manually and link data for research via a master population list for a designated region. Recently, a more flexible computerized third-party linkage platform, Secure Decoupled Linkage (SDLink), has been proposed based on: (1) decoupling data via encryption, (2) obfuscation via chaffing (adding fake data) and universe manipulation; and (3) minimum information disclosure via recoding.

Results We synthesize this literature to formalize a new framework for privacy preserving interactive record linkage (PPIRL) with tractable privacy and utility properties and then analyze the literature using this framework.

Conclusions Human-based third-party linkage centers for privacy preserving record linkage are the accepted norm internationally. We find that a computer-based third-party platform that can precisely control the information disclosed at the micro level and allow frequent human interaction during the linkage process, is an effective human-machine hybrid system that significantly improves on the linkage center model both in terms of privacy and utility.

INTRODUCTION

Information systems in the health sector have undergone significant infrastructure changes making it possible to collect, store, and process huge amounts of data. However, information derived from these heterogeneous systems is often redundant, fragmented over multiple databases, incomplete, and erroneous.^{1–7} In fact, the 4V's of *Big Data*, Volume, Velocity, Variety, and Veracity,⁸ describe succinctly the nature of *Big Data* in health-care as seen in the continuously generated medical records from diverse service providers which always contain some level of error. Thus, a task critical to finding the useful information among such chaotic *Big Data* is record linkage—the process of identifying record pairs from different information systems which belong to the same real-world entity.

The record linkage process is complicated by the inherent factors observed in *Big Data*, such as missing data (eg, missing social security number (SSN)), erroneous data (eg, transpose of date of birth (DOB)), non-standardized forms of data (eg, Dr Smith), and change in the data over time (eg, changed last name). The absence of common, error-free, and unique identifiers makes exact matching solutions inadequate, leading to methods for approximate linkage to address these issues.^{1–15} In a study linking cancer registries, 10% more matches were found using a deterministic approximate match compared to the exact match methods due to typos in names or missing SSNs.² A more sophisticated approximate method, six pass probabilistic record linkage, linking a cancer registry with Medicaid data, reported only 83% of records were matched using exact match.³ In another study, 36.3% of health records were missing SSNs.⁵ Yet another study reported that there were between 0.16% and 16% potential duplicate medical record numbers in five different electronic health record systems.⁶ Due to the large number of patients served, even 0.16% equals 1583 records, quite a considerable number to clean up manually.

In this paper, we provide a tutorial on record linkage and a systematic review of the literature on privacy preserving record linkage (PPRL) for biomedical research. We also synthesize the literature to propose a new framework, privacy preserving interactive record linkage (PPIRL), for data integration with tractable privacy and utility properties. We evaluate the current literature using the framework.

BACKGROUND

Record linkage

The main difficulty in record linkage is that data are often expressed differently, change over time, lack unique attributes, have missing attributes, or have erroneous data entry. Let us consider an example where SSN, first name, last name, and DOB are available for linkage. If we link only on SSN, issues arise from missing and erroneous SSN. If linked using all four attributes on exact match, many true matches are missed. The goal of the different approximate approaches is to capture as many true matches as possible while minimizing the false matches. Typically, all approaches will use approximate matching and result in three categories: match, uncertain, non-match. The objective in all automatic approximate algorithms is to minimize the uncertain region which requires manual resolution by an individual. There are several good surveys^{9–15} and recent advances in new learning methods for automatic matching.^{16–21}

To cite: Kum H-C, Krishnamurthy A, Machanavajjhala A, et al. *J Am Med Inform Assoc* 2014;**21**:212–220.

Uncertainty in record linkage

In all approximate matching methods, real-world entities which share similar identifying information (eg, twins and family) result in a certain number of false matches.^{3 5 22} In addition, in health informatics application, there are substantial ethical and liability issues involved in the potential corruption of the integrated patient data system that can result from these false matches.^{6–7 23 24} At the same time, conservative matching methods which can miss many true matches, may result in selection biases.^{4–5 25} Not properly accounting for linkage error, both false matches and missed true matches, can cause serious harm as the erroneous links propagate to subsequent steps in the workflow.^{6 7 22–30} Bronstein *et al*⁴ found that when matching Medicaid claims data to vital records, the resulting matched analytic datasets tend to under-represent the outcomes of high-risk pregnancies. Baldi *et al* found that the covariates in the Cox regression models can be biased due to not capturing all true links when analyzing survival rate in a cohort of patients with breast cancer.²⁵ Lahiri and Larsen propose a method for taking into account the measurement error in the linkage process when building a linear regression model between linked variables.²⁶ Tancredi and Liseo present a more general model for propagating the uncertainty between the parameter estimation step and the matching procedure using a hierarchical Bayesian approach.²⁷

Currently, most research treats linked data as if there are no errors. This convention is perpetuated because most scientists using the linked data are not involved with the linkage process⁷ and do not fully appreciate the complex process or the uncertainties in the linked data. Researchers who use linked data need a better understanding of the nature of uncertainty in the linkage process and more research is needed on methods of propagating the uncertainty in record linkage to subsequent analysis.^{25–29}

Interactive record linkage

Linkage errors propagate into the linked data and its analysis results leading to potential problems with incorrect results, and eventually incorrect knowledge and action. Thus, *interactive record linkage*, defined as people fine tuning the false matches and managing the uncertainty and its propagation to subsequent analyses, is the first step in the data workflow to turn *Big Data* into useful biomedical information.^{3–5 31} We define the properly tuned output from such a hybrid human-machine data integration system as *high quality record linkage*.

Recently, there has been more research on interactive record linkage that takes advantage of human interaction either through active learning systems or crowdsourced systems^{32–38} after a study described the limitations of the techniques in automatic record linkage for real applications.³⁹ More research is needed on interactive record linkage systems that allow the scientist to tune the linkage results and manage the uncertainty in the subsequent analysis. The importance of human interaction in record linkage to resolve the many uncertainties in the process is demonstrated well in Bronstein *et al*.⁴ Their paper describes a method for matching pregnancies from Medicaid data to birth records using probabilistic record linkage that involved 11 manual steps. There were multiple uncertainties that needed human decisions during the process. For example in step 4, of 46 364 pregnancies the authors were trying to match, 4369 linked to more than one vital record and 9400 had no match to any vital record. Eventually after multiple iterative data cleaning and matching steps, the authors identified 43 500 completed pregnancies that should be documented in vital

records, 5278 of which were not found (87.9% match rate). This is similar to the 90% match rates found in linking medicaid and vital statistics records in other states in the USA. With no human interaction, the match rate would be much lower. Such a high level of human interaction and iteration is common in medical record linkage studies.^{3–5}

Privacy in record linkage

Given the sensitivity of biomedical data, privacy is a major concern in interactive record linkage where data cannot be de-identified. In particular, in secondary data analysis the research question is not known at the time of data collection, making informed consent, the most common form of protection in biomedical research, difficult. In most cases, general blanket consent for research along with IRB review of the risks and benefits of research, is the only option available. In 2001, the US Government Accountability Office (GAO) published a report on technologies for privacy protection in record linkage in federally funded projects.⁴⁰ Much is still the same with only two modes of access for research, de-identify mode and trust mode. De-identified data cannot be linked and the trust mode provides little protection from trusted users requiring high level clearance for those doing the linkage. In addition, with trusted third-party linkages, scientists are unaware of the uncertainty in the linkage process and how to propagate this uncertainty in the analysis downstream.²²

Privacy protection in record linkage is fundamentally different from all other privacy preserving data operations^{41 42} because the goal is to exactly identify the entity represented by the data being linked, so that the tables can be accurately merged.²² Thus, there is a direct conflict with the conventional understanding of privacy as anonymity. More precisely, anonymity is preventing *identity disclosure*. In comparison, *attribute disclosure* refers to the disclosure of one or more sensitive attributes (eg, cancer status). Although related, identity disclosure is only a necessary condition for attribute disclosure, not a sufficient condition. Identity disclosure without attribute disclosure has a low risk of harm.^{22 43–47} For safe interactive record linkage, we need to find the exact level of information disclosure that protects sensitive data but reveals enough identifying data for high quality linkage.

Thus, a model focusing on guaranteeing no attribute disclosure while also minimizing identity disclosure has the potential to significantly reduce the risk of privacy violation while still allowing for high quality data integration.^{46 47} It is important to note that the current norm for data integration in the USA is full disclosure of all information to a fully trusted human entity, often called the honest broker; for example, full disclosure of both attribute and identity to certain trusted parties for certain purposes is HIPAA (Health Insurance Portability and Accountability Act) compliant.^{2–7} Often, the trusted party is a government or hospital employee, or business associates who must access identifying information for operations.⁷ Typically in biomedical research, the trusted party is responsible for maintaining a master patient index (MPI), and this index is used to integrate data. The quality of MPI varies widely and most MPIs have duplicates that must be cleaned during the linkage process.⁶

In this current trust model, there is no protection from insider attack. The main threat model in the interactive linkage process is an honest-but-curious (HBC) user who follows protocol⁴⁸ but carries out an insider attack, which accounts for close to half the breaches in the USA.⁴⁹ In a survey of over 600 people, 46% of the respondents answered that the damage

caused by insider attacks is greater than that caused by outsider attacks, with the most common insider e-crimes being unauthorized access to and use of information (63%) and unintentional exposure of private and sensitive data (57%).⁴⁹ By focusing on giving trusted parties access to only the minimum information required, unintentional exposure of sensitive data can be significantly reduced.

PRIVACY PRESERVING RECORD LINKAGE

For our systematic review of the topic, we modified the guidelines of the Center for Reviews and Dissemination.⁵⁰ Figure 1 details the workflow with specifics provided in the online supplementary appendix. Here we present the three themes that emerged from the 71 articles reviewed in two separate sections, 'Privacy preserving record linkage' and 'Privacy preserving interactive record linkage.'

Private record linkage

On the theoretical front, there have been ongoing efforts to develop PPRL algorithms since 2003.⁵¹ Private record linkage is defined as computing the set of linked records given as input a matching function and then outputting them to the two private parties without revealing anything about the non-linked records. The first generation of private record linkage algorithms relied on hash-based algorithms.^{51–52} The use of hashes resulted in strong privacy guarantees but was limited to exact matching algorithms. This led to the second generation of algorithms that developed private string comparison methods (eg, Bloom filters) for private approximate matching.^{53–60} Secure multi-party computation (SMC) is also a common approach to protect against cryptographic attack. Several surveys of private record linkage^{61–64} and privacy-preserving string comparators⁶⁵ have been carried out. Recently, Kuzu *et al* proposed a practical private record linkage system demonstrating the effectiveness of controlled information disclosure via obfuscated data and SMC.^{66–67} However, they still formulate the problem as private record linkage with a known matching function and ambiguous links.

In summary, private record linkage involves two private parties who are trying to share minimum information with each other and assumes that the matching function between the tables is known. The goal is to apply the known matching function in a secure manner. There are two problems in this formulation. First, if the matching function is not known, as in most applications, the algorithms cannot be used. Second, there is no possibility of clerical review of the ambiguous links or human interaction during linkage because one of the assumptions is that the private data must not be revealed to the other party.

Consequently, the major challenges for real applications are that, without human interaction, there is no method for finding the matching function and resolving the ambiguous links. All reviews of private record linkage identify these as open issues.^{61–64}

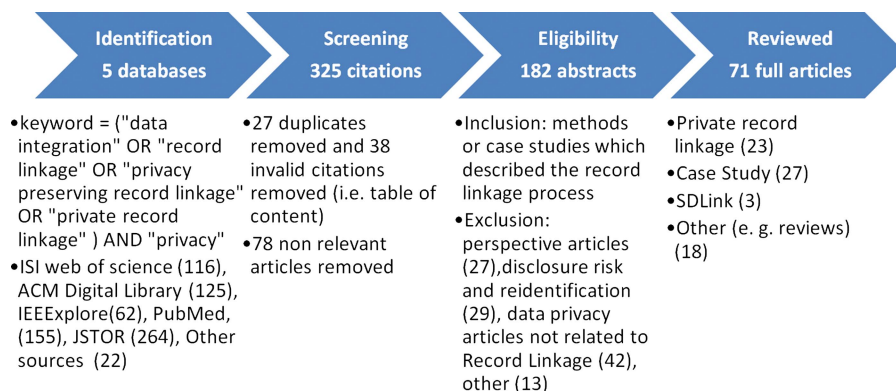
Trusted third-party linkages

In practice, published research using linked data uses a trusted third-party model.^{2–7} In the USA, the National Center for Health Statistics or state centers for health statistics often play the role of a trusted third party. Internationally, several countries have linkage centers whose main role is to determine the matching function manually and link data for research as the trusted third party. Many linkage centers have succeeded in building systems for integrating population health records with good protocols for privacy protection,^{5–68–76} sometimes called the pseudonym approach. Such centers rely on separation of the identifying information from the sensitive information for privacy protection.^{5–75–78} Dedicated record linkage experts have access to only the identifying data with no access to sensitive data, and furthermore are not involved with subsequent research using the linked data. In these linkage centers, there is significant reliance on the human expert for high quality record linkage and maintenance of a master population list to which all data are linked. Hertzman *et al*⁷⁵ describe this proactive linkage as 'linking each data set when it arrives from a data provider, rather than project by project.' Most linkage centers cover a designated region, easing the burden of maintaining a master population list, and operate in countries with uniformity in health records and national identification numbers.

PRIVACY PRESERVING INTERACTIVE RECORD LINKAGE

In a heterogeneous health system, like that in the USA, the validity and reliability of integrated health data is a significant problem.^{5–7–22–24} In these settings, given the velocity and veracity of *Big Data*, good incremental record linkage methods are required for proactive linkage to work well since multiple data continue to flow into the system with no shared unique identifiers. However, incremental record linkage to maintain a coordinated master list and its links to multiple data sources that change over time is still largely an open research area. The literature confirms that high quality data integration as well as managing uncertainty in *Big Data* require human interaction throughout the entire workflow.^{1–7–22–30} Human interaction means that data must be revealed to someone in some form under some condition. In this section, we synthesize the literature to propose PPIRL, a novel framework with tractable privacy and utility properties. We then review a system called

Figure 1 Systematic review process workflow.



Secure Decoupled Linkage (SDLink) as a possible implementation of the principles of PPIRL.

The main use case for the PPIRL framework is for those with approval for full access to multiple data under the trust mode for data integration. PPIRL is a framework that can protect against HBC users in such situations. If successful, PPIRL can greatly increase the throughput of record linkage by allowing many more people, who have *not* obtained the highest trusted party status (eg, graduate students), to be involved in the time consuming steps of record linkage, creating the matching function, and carrying out the clerical review. Furthermore, under certain conditions, crowdsourcing parts of the process is possible.

The goals of PPIRL are to allow direct control of the matching function and the matching uncertainty by the user while still providing privacy protection, defined as no sensitive attribute disclosure, during this interaction. In box 1, we present the PPIRL framework.

The cost of privacy in PPIRL is the difference in quality of the matching functions M' and M . The key to solving the PPIRL problem is to understand the minimum amount of information required for the human user, H , to make accurate linkage decisions and then to devise methods to disclose that information to H without disclosing any of the sensitive attributes S_1 and S_2 from the databases R_1 and R_2 . If we can disclose all of the information required for generating the matching function M safely, then the quality of the matching function M' can be as good as M and privacy can be guaranteed at no cost to utility.

Box 1 Privacy preserving interactive record linkage (PPIRL)

Problem statement (interactive record linkage, IRL) Let R_1 and R_2 be two private datasets, which cover data on subsets of a population Ω , with non-sensitive attributes Q_1 and Q_2 , and sensitive attributes S_1 and S_2 , respectively. The goal of IRL is to construct an algorithm A that takes as input R_1 and R_2 , and outputs a function $M: R_1 \times R_2 \rightarrow \{\text{match, non-match}\}$ AND a function $C_M: R_1 \times R_2 \rightarrow [0,1]$. The function C_M is an automatic function which outputs a probability score of match between 0 and 1 reflecting the confidence level of the match. The function M is also automatically computed, but for selected mappings, typically from uncertain regions, the output assignment can be interactively changed by an informed human H . In IRL, the human H has access to the full datasets R_1 and R_2 , as well as the output from C_M to tune the final matching function M .

Problem statement (privacy preserving interactive record linkage, PPIRL) The goal of PPIRL is to construct an algorithm, A' that outputs function M' and $C_{M'}$, which serves the same purpose as algorithm A from IRL except that the sensitive attributes S_1 and S_2 , from the datasets R_1 and R_2 , respectively, are not disclosed to the human H . The human in PPIRL is thus typically working with less data about the records being linked but trying to still achieve the same level of quality in the matching function M' .

Privacy objective To protect against sensitive attribute disclosure, S_1 and S_2 are never revealed to H .

Utility objective (1) To generate the best matching function M' possible by allowing a person H to fine tune the results; and (2) to generate and communicate the confidence level $C_{M'}$ to H , so that uncertainty can be managed and propagated through the full analysis workflow flexibly.

Secure decoupled linkage

SDLink is a flexible, secure linkage system that implements the key ideas behind PPIRL. Below we review the key privacy design principles of the system.^{22 46 47}

Privacy design 1: decoupling data

Decoupling refers to separating out, via encryption, the identifying information (eg, personally identifiable information (PII)) from the sensitive data (eg, cancer status) that need protection (figure 2).^{46 47} Decoupled data provide the same level of protection as de-identified data, but with more protection than is provided in the trust mode of access and more utility.^{22 46 47} Decoupling data follows the minimum necessary standard for privacy protection and, during the record linkage process, removes unnecessary information, that is, the information connecting the PII to sensitive data. The innovation in decoupling data is to take a privacy-by-design approach and focus on selectively revealing information rather than hiding it. The key is to understand the minimum information required for acceptable linkage and then to design protocols to reveal, in a secure manner, only that information.

Privacy design 2: computerized third-party linkage

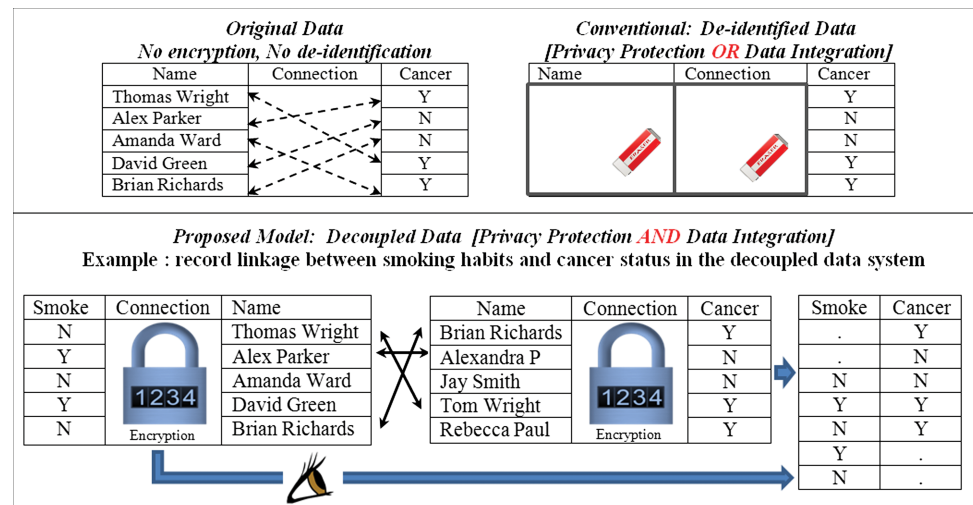
As discussed above, the trusted third-party mechanism to protect privacy is well understood. In the decoupled approach, a researcher has access to computerized third-party software that can access the PII in order to link the data. The researcher requests that two decoupled tables be merged, after which the computerized third-party software takes control and carries out the linkage. In this process, the software actively interacts with the researcher as needed for guidance on parameter settings (determining the matching function) and resolving ambiguities (clerical review of ambiguous links). Essentially, a decoupled data access system is a computerized third-party equivalent to the human-run data linkage centers that strictly controls information.

The main benefit of a computerized third-party model of privacy protection is that it allows each project to have maximum flexibility in its linkage to control the uncertainty in real data. With properly designed third-party software acting as an oracle, a person can interact frequently and inexpensively with information held by the computer third party at the smallest level (eg, asking how similar two encrypted SSNs are) in order to manage uncertainty in the linkage. The decoupled database software functions like a bank vault with security deposit boxes that have well-developed security protocols for importing and accessing datasets in the system. The system only allows access to particular tables to those who have the appropriate decryption keys which are managed by the different data custodians.

Privacy design 3: chaffing and universe manipulation

With decoupling, researchers cannot associate a particular row of data with any PII disclosed during record linkage. But researchers can combine what they know with the PII data shown during interaction to make certain inferences and learn sensitive information about people they know.^{46 47} The privacy literature has shown that background knowledge can be used to infer more information than is originally disclosed.^{44 45} For example, in homogeneous data, attribute disclosure can occur via group membership⁴³ (eg, someone you know is on the list of cancer patients). Thus, strict decoupling via encryption is not sufficient to protect against attribute disclosure when identities could be revealed during human interaction. The probability of

Figure 2 Secure decoupled data. Internally, the data is stored in a decoupled data system (bottom), which has the same level of privacy protection as de-identified data (top right), but is much more powerful because researchers can link multiple decoupled datasets safely. Decoupled data allows for accurate record linkage with no attribute disclosure.



attribute disclosure through group membership is dependent on a variety of factors including any pre-existing information that is known by the observer, the knowledge of the nature of the list, and the uniqueness of the PII in the universe of the data.⁴⁶⁻⁴⁷ To overcome this, Kum *et al*⁴⁶⁻⁴⁷ evaluated three methods of modification (figure 3): (1) *chaffing*: literally changing the nature of the universe by adding fake data; (2) *fabrication*: changing the label/name of the universe presented to mislead the user about the nature of the list; and (3) *non-disclosure*: hiding

the identity of the universe to reduce confidence by making the list less tractable (eg, a list from the USA compared to a list from Austin). The study showed that when the universe around the data was not disclosed, 56% of the participants were uncertain about the identity given a common name. Even for rare names, if the list is chaffed and the universe is not disclosed, 66% of the participants were uncertain about the identity.²² These results show that through chaffing and universe manipulation, identity disclosure can be minimized for both common and rare names.

Figure 3 Chaffing and Universe Manipulation. Triangles: cancer patients; cross-hatched circles: not cancer patients. D_A : Universe of all cancer patients (eg, USA); L_A : list of subset of cancer patients being reviewed for linkage which is more tractable (eg, Austin); lan_{PII} represents the PII of someone that the reviewer knows (eg, Ian who lives in Austin). Since Ian is not a unique name, it is unclear whether the PII represents the same real world Ian that the reviewer knows personally. (1) *chaffing*: literally changing the nature of the universe by adding fake data (eg, add blue circles to red triangles); (2) *fabrication*: changing the label/name of the universe presented to mislead the user on the nature of the list (eg, label D_A as D_B and/or L_A as L_B , thus lan_{PII} now is presented as someone who lives in Beijing, China, who could not be the same Ian that the reviewer knows to live in Austin); and (3) *nondisclosure*: hiding the identity of the universe to reduce confidence by making the list less tractable. That is, by not disclosing the label L_A , the user must assume the list represents a much larger universe D_A (eg, a list from USA compared to list from Austin). The reviewer, who knows an Ian living in Austin, loses confidence in inferring the real identity of lan_{PII} when it is presented as an Ian living in the USA compared to being presented as an Ian living in Austin.

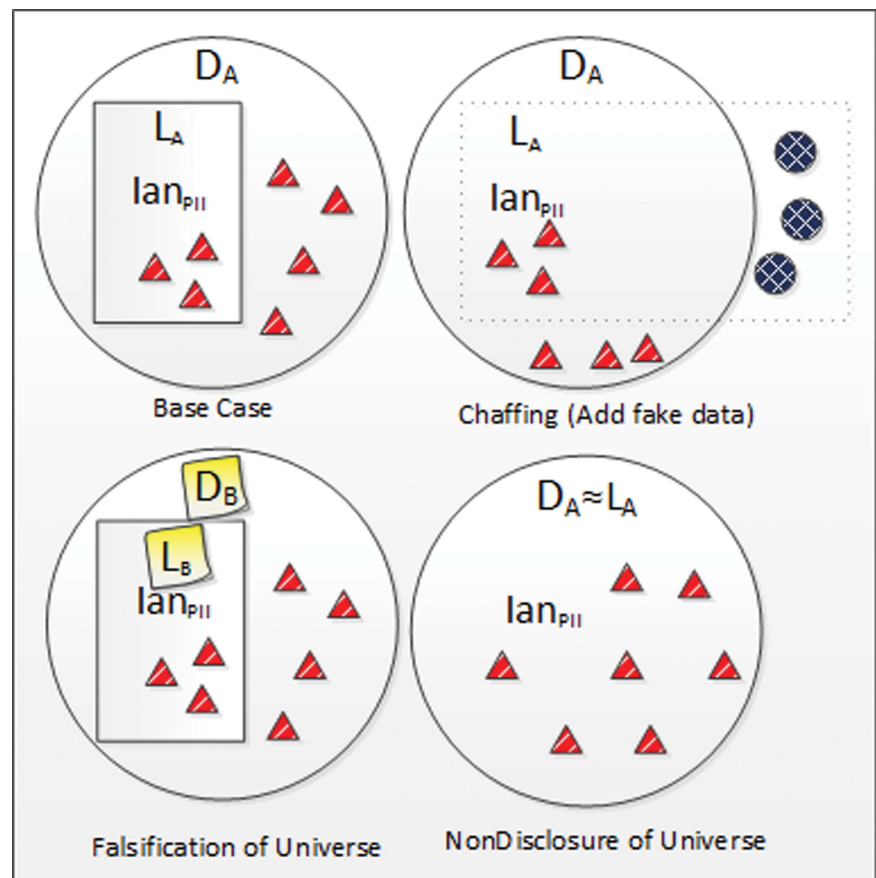








Figure 4 Data recoding techniques.⁴⁷

The SDLink GUI applies data recoding techniques which display the difference between the attributes that are meaningful for record linkage instead of the raw data. For example, the gender field only indicates, same[–], different[D], or missing[M] in one or both fields. DOB, date of birth; SSN, social security number.

D=Different M=Missing TX=Transpose —=Same
Type Y/N in the right most column for linkage

Rec. No.	First Name	Last Name	DOB mm/dd/yyyy	SSN	Gen der	Link (Y/N)
111	John	Gray	--/D-/--D-	--D--D---	M	
	Jon	Grey				
112	Alex	Parker II	T/X/---	-----	-	
	Alex	Parker				
113	Donna	Balmer	--/--/---	----TX--	-	
	Donna	Palmer				
114	Timothy	Richards	--/TX/---	M	-	
	Mr. Tom	Richards				
115	Anita	Gorge	--/--/--TX	--D--D---	D	
	Anita	George				
116	Michael	Smith	--/--/---	-----	M	
	Michael	S				

Privacy design 4: minimum information sharing via recoding

What information is disclosed during the interaction with the decoupled system determines the risk of disclosure.^{43–47}

Figure 4 depicts a sample screen during the linkage process with only the name being fully disclosed. The differences between the attributes that are meaningful for record linkage are displayed instead of the raw data.^{46–47} For example, the gender field only indicates same, different, or missing in one or both fields. Identity (ID) numbers which are PII with a risk of harm (eg, SSN), are displayed as the number of different digits and transposes. DOB comparisons are made on an element basis for month, day, and year. In addition, transpose of month and day is accounted for as well as transposes within one element. Determining meaningful differences in names is the most difficult. Table 1 depicts different levels for data recoding of names from left (high disclosure) to right (low disclosure). More research is required to understand what level of information will result in acceptable levels of high quality linkages from interactive record linkage.

Comparing PPiRL with existing data linkage methods

Although many private record linkage systems have strong privacy guarantees, it is assumed that the matching function is known and thus has a different objective than PPiRL. The use case for PPiRL is similar to that in the linkage centers where there is one trusted party with access to all the required data for linkage. The trusted third-party model can, to some extent, meet the privacy objective of PPiRL if the sensitive data are isolated from the identifying data. However, better documentation is required on the detailed protocols to handle threats by the HBC trusted users who can disclose information unintentionally and/or access unauthorized data. As discussed above, separation alone will not guarantee that no attribute is disclosed due to homogenous group membership. On the other hand, the SDLink platform can guarantee no sensitive attribute disclosure by: (1) decoupling sensitive data from identifying data via encryption; (2) using chaffing to block against attribute disclosure via group membership along with universe manipulation; and (3) recoding to minimize identity disclosure.

Table 1 Different data recoding techniques for names

Record no.	Full disclosure	Remove identical strings	Edit distance if small	Edit distance	Length:edit distance and frequency		Binary
111	Gray	Gray	–a–	–a–	4:1	Rare	DIFF
	Grey	Grey	–e–	–e–	4:1	Common	DIFF
112	Parker II	— II	— II	— II	7:1	Common	DIFF
	Parker	—	—	—	6:1	Common	DIFF
113	Balmer	Balmer	B—	B—	6:1	Common	DIFF
	Palmer	Palmer	P—	P—	6:1	Common	DIFF
114	Richards	—	—	—	0	Very common	SAME
	Richards	—	—	—	0	Very common	SAME
115	Carey	Carey	Carey	–ey	5:2	Common	DIFF
	Carr	Carr	Carr	–r	4:2	Common	DIFF
116	Smith	Smith	Smith	–mith	5:4	Very common	DIFF
	S	S	S	–	1:4	Rare	DIFF

The first utility objective of generating the best possible matching function by a human expert is met in the linkage centers. However, the objective at the linkage centers is to maintain a global optimal matching function for all data at all times, which can be difficult in many circumstances involving continuously changing heterogeneous data. In comparison, the SDLink platform has been built to allow for finding a local optimal matching function on a per project basis depending on the data required to be merged for a given study. In reality, developing totally new matching functions for every project will be too expensive. But, the ability to optimize existing matching functions per project will allow for better tuning of linkage results and less uncertainty because the scope of the problem per project is significantly smaller than the global problem. Although PPIRL only discusses one matching function M' for simplicity, in reality there are multiple matching functions that meet the diverse needs of different projects.⁷⁹ A flexible system that can efficiently support multiple matching functions is required to give the scientists the control they need over the data to propagate and manage the uncertainty of *Big Data*. The SDLink platform is a safe infrastructure that can be utilized by many scientists to carry out all aspects of record linkage research including a model for uncertainty propagation while protecting privacy.

Discussion of PPIRL and SDLink

To the best of our knowledge, the research on SDLink platform is based on good privacy designs, reviewed in detail in this paper, and best meets both the privacy and utility objectives of PPIRL. SDLink proposes a platform to improve on the existing record linkage centers in terms of both privacy and utility. Nonetheless, it is unclear how effective the privacy designs proposed will fare against more vicious adversaries with background information. More research is required on precisely what information a person needs for tuning the linkage results and the harm that can result from release of just that information given the readily available background information in the digital age. Any wide table with many attributes required for biomedical research when combined with publically available background information may release more information than is safe even if it is de-identified. Thus, along with privacy research that guarantees the minimum release of information required for biomedical research, better research is needed on how to make sure that the information released is properly protected. The strongest confidentiality protection is provided by secure data centers that strictly control physical access to the data by not allowing remote access. However, the various costs associated with such data centers are prohibitive. Thus, a data infrastructure based on a more holistic coordinated approach that combines methods from technology, statistics, policy, and ethics is required so that *Big Data* can be used for biomedical research.^{5 7 22 31 80} An extensible platform for building a comprehensive knowledge base that meets the needs of biomedical research is quite complex and managing digital entities is at the core of the problem. Bellare *et al* present a good starting point for continuously maintaining huge numbers of digital entities for a continuous knowledge base in the context of search engines⁷⁹ that should be extended with privacy guarantees for biomedical research.

CONCLUSIONS AND FUTURE WORK

Privacy preserving data integration is key to any data intensive biomedical research using *Big Data*. Given the volume, variety, velocity, and veracity of *Big Data*, tuning the results of

automatic record linkage algorithms via human interaction is the only way to achieve high quality record linkage as well as manage and propagate the uncertainty in the linked data. A properly designed computerized third-party platform, such as SDLink, that can precisely control the information disclosed at the micro level and allows frequent human interaction during the linkage process, is an effective human-machine hybrid system that can accurately and safely integrate *Big Data* for biomedical research.

Sometimes the quality of linkage can be improved when sensitive data are available during linkage. For example, sorting through twin records is easier done with sensitive data. The right trade-off between the quality of linkage and protection must be case dependent and should be determined by an IRB based on the risk and benefit, considering issues such as who are doing the linkage, on what computer system and with what software, and for what purpose. Most importantly, for population level research, as long as there are means to propagate and bound errors from linkage the optimal linkage may not be required. More research is needed on: (1) precisely what information is required for good linkage decisions; (2) how to disclose only that information in an effective and safe manner; (3) possible threats from and countermeasures against more aggressive adversaries; and (4) how to propagate the uncertainty in record linkage to subsequent analysis steps.

Acknowledgements We thank Darshana Pathak and Jan Werner for their helpful comments.

Contributors All co-authors worked together on this research and wrote the manuscript.

Funding None.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

- 1 Sauleau EA, Paumier J, Buemi A. Medical record linkage in health information systems by approximate string matching and clustering. *BMC Med Inform Decis Mak* 2005;5:32–44.
- 2 Weber SC, Lowe H, Das A, *et al*. A simple heuristic for blindfolded record linkage. *J Am Med Inform Assoc* 2012;19:157–61.
- 3 Boscoe FP, Schrag D, Chen K, *et al*. Building capacity to assess cancer care in the Medicaid population in New York State. *Health Serv Res* 2011;46:805–20.
- 4 Bronstein J, Lomatsch C, Fletcher D, *et al*. Issues and biases in matching Medicaid pregnancy episodes to vital records data: the Arkansas experience. *Mater Child Health J* 2009;13:250–9.
- 5 Duvall SL, Fraser AM, Rowe K, *et al*. Evaluation of record linkage between a large healthcare provider and the Utah population database. *J Am Med Inform Assoc* 2012;19:e54–9.
- 6 McCoy AB, Wright A, Kahn M, *et al*. Matching identifiers in electronic health records: implications for duplicate records and patient safety. *BMJ Qual Saf* 2013;22:219–24.
- 7 Bradley C, Penberthy L, Devers K, *et al*. Health services research and Data Linkages: issues, methods, and directions for the future. *Health Serv Res* 2010;45 (5p2):1468–88.
- 8 Kash W. Federal Lab, IBM Open Door To High Speed Computing For Big Data Users. Website June 28, 2012. <http://breakinggov.com/2012/06/28/federal-lab-ibm-open-door-to-high-speed-computing-for-big-data/>
- 9 Elmagarmid K, Panagiotis GI, Verykios SV. Duplicate record detection: a survey. *IEEE Trans Knowl Data Eng* 2007;19:1–16.
- 10 Winkler WE. Overview of record linkage and current research direction. Research Report Series (Statistics No. 2006-2), Statistical Research Division, U.S. Census Bureau. Washington, DC, 2006.
- 11 Herzog T, Scheuren F, Winkler W. *Data quality and record linkage techniques*. Springer, 2007.
- 12 Christen P. *Data matching*. Springer, 2012.
- 13 Getoor L, Machanavajjhala A. Entity resolution: theory, practice & open challenges. *International Conference on Very Large Data Bases (VLDB)*. 2012.

- 14 Newcombe H, Kennedy J, Axford S, *et al.* Automatic linkage of vital records. *Science* 1959;130:954–59.
- 15 Fellegi P, Sunter AB. A theory for record linkage. *JASA* 1969;64:1183–210.
- 16 Benjelloun O, Garcia-Molina H, Menestrina D, *et al.* Swoosh: a generic approach to Entity Resolution. *VLDBJ* 2009;18:255–76.
- 17 Christen P. Automatic record linkage using seeded nearest neighbor and support vector machine classification. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08). New York, NY, USA: ACM, 2008:151–9. doi:10.1145/1401890.1401913
- 18 Bhattacharya I, Getoor L. A latent Dirichlet model for unsupervised entity resolution. *SDM* 2007.
- 19 Bilenko M, Mooney R. Adaptive duplicate detection using learnable string similarity measures. *KDD* 2003.
- 20 Chen Z, Kalashnikov DV, Mehrotra S, *et al.* Exploiting context analysis for combining multiple entity resolution systems. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD '09). Carsten Binnig and Benoit Dageville, eds. New York, NY, USA: ACM, 2009:207–18. doi:10.1145/1559845.1559869
- 21 Sadinle M, Hall R, Fienberg SE. Approaches to multiple record linkage. *Proceedings of ISI*. Dublin, Ireland: 2011.
- 22 Kum HC, Ahalt S, Pathak D, Privacy preserving data integration using decoupled data. In: Elovici Y, Altschuler Y, Cremers A, Aharoni N, Pentland A. eds. *Security and privacy in social network*. Springer, 2012:225–53.
- 23 Boyd A, Hosner C, Hunscher D, *et al.* An "Honest Broker" mechanism to maintain privacy for patient care and academic medical research. *Int J Med Inform* 2007;76:407–11.
- 24 Boyd A, Saxman P, Hunscher D, *et al.* The University of Michigan honest broker: a web-based service for clinical and translational research and practice. *J Am Med Inform Assoc* 2009;16:784–91.
- 25 Baldi I, Ponti A, Zanetti R, *et al.* The impact of record-linkage bias in the Cox model. *J Eval Clin Pract* 2010;16:92–6.
- 26 Lahiri P, Larsen M. Regression analysis with linked data. *J Am Stat Assoc* 2005;100:222–30.
- 27 Tancredi A, Liseo B. A hierarchical Bayesian approach to record linkage and population size problems. *Annu Appl Stat* 2011;5:1553–85.
- 28 Scheuren F, Winkler WE. Regression analysis of data files that are computer matched—Part II. *Surv Methodol* 1997;23:157–65.
- 29 Scheuren F, Winkler WE. Regression analysis of data files that are computer matched. *Surv Methodol* 1993;19:39–58.
- 30 Fienberg S. Toward a reconceptualization of confidentiality protection in the context of linkages with administrative records. *J Privacy Confidentiality* 2011;3:65–71.
- 31 Kum HC, Ahalt S. Privacy by design: understanding data access models for secondary data. American Medical Informatics Association (AMIA) Joint Summits on Translation Science and Clinical Research Informatics. 2013. Nominated for best paper award.
- 32 Wang J, Kraska T, Franklin MJ, *et al.* CrowdER: Crowdsourcing Entity Resolution. *Proceedings of Very Large Data Bases (VLDB)* 2012;5.
- 33 Kang H, Getoor L, Shneiderman B, *et al.* Interactive entity resolution in relational data: a visual analytic tool and its evaluation. *IEEE Trans Vis Comput Graph* 2008;14:999–1014.
- 34 Marcus A, Wu E, Karger D, *et al.* Human-powered sorts and joins. *Proceedings of Very Large Data Bases (VLDB)* 2011;5.
- 35 Arasu A, Gotz M, Kaushik R. On active learning of record matching packages. *ACM International Conference on Management of Data (SIGMOD)*. 2010.
- 36 Beygelzimer A, Langford J, Hsu D, *et al.* Agnostic active learning without constraints. *NIPS*, 2010. <http://arxiv.org/abs/1006.2588>
- 37 Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning. *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*. 2000.
- 38 Bellar K, Iyengar S, Parameswaran A. Active sampling for entity matching. *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*. 2012.
- 39 Kopcke H, Thor A, Rahm E. Evaluation of entity resolution approaches on real-world match problems. *PVLDB* 2010;3:484–93.
- 40 GAO. Record linkage and privacy: issues in creating new federal research and statistical information, GAO-01-126SP, April 2001 (172 pp).
- 41 Vaidya J, Zhu Y, Clifton C. *Privacy preserving data mining*. Advances in information security. New York: Springer-Verlag, 2005.
- 42 Jiang X, Sarwate A, Ohno-Machado L. Privacy technology to support data sharing for comparative effectiveness research: a systematic review. *Med Care* 2013; 51(8 Suppl 3):S58–65.
- 43 Fienberg SE. *Confidentiality, privacy and disclosure limitation*, *Encyclopedia of Social Measurement*. Academic Press, 2005;1:463–9.
- 44 Machanavajjhala A, Kifer D, Gehrke J, *et al.* L-diversity: privacy beyond *k*-anonymity. *ACM Trans. Knowl Discov Data* 1, 1, Article 3 (March 2007).
- 45 Li N, Li T, Venkatasubramanian S. *t*-Closeness: privacy beyond *k*-anonymity and *l*-diversity. *International Conference on Data Engineering (ICDE)*, April 2007.
- 46 Kum HC, Pathak D, Sanka G, *et al.* Privacy beyond anonymity: decoupling data through encryption for record linkage. Technical Report 2012-003 UNC-CH. Poster Presentation at American Medical Informatics Association (AMIA) Joint Summits on Translation Science and Clinical Research Informatics. 2013.
- 47 Kum HC, Pathak D, Krishnamurthy A, *et al.* Secure Decoupled Linkage (SDLink) for Building a Social Genome. In: Proceedings of 2013 IEEE International Conference on Big Data (IEEE BigData 2013), Oct 2013. San Jose, CA. 7–11.
- 48 Goldreich O. Secure multi-party computation. Unpublished manuscript. Final Version. October 2002. <http://www.wisdom.weizmann.ac.il/~oded/pp.html>
- 49 2011 CyberSecurityWatch Survey, CSO Magazine, U.S. Secret Service, Software Engineering Institute CERT Program at Carnegie Mellon University and Deloitte, January 2011.
- 50 Tacconelli E. Systematic reviews: CRD's guidance for undertaking reviews in health care. *Lancet Infect Dis* 2010;10:1–232.
- 51 Agrawal R, Evfimievski A, Srikant R. Information sharing across private databases. *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*. New York, USA: ACM, 2003:86–97.
- 52 Freedman MJ, Nissim K, Pinkas B. Efficient private matching and set intersection. *Proceedings of EUROCRYPT*. 2004.
- 53 Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. *BMC Med Inform Decis Mak* 2009;9:41.
- 54 Churches T, Christen P. Some methods for blindfolded record linkage. *BMC Med Inform Decis Mak* 2004;4:9.
- 55 Yakout M, Atallah MJ, Elmagarmid AK. Efficient private record linkage. In *ICDE*. 2009:1283–6.
- 56 Durham E, Xue Y, Kantarcioglu M, *et al.* Private medical record linkage with approximate matching. *Proceedings of the 2010 American Medical Informatics Association Annual Symposium*. 2010:182–6.
- 57 Schroeder AD. Pad and Chaff: secure approximate string matching in private record linkage. *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services (IIWAS '12)*. New York, USA: ACM, 2012:121–5.
- 58 Karakasidis, Alexandros, Verykios, *et al.* Fake injection strategies for private phonetic matching. In: Garcia-Alfaro J, Navarro-Arribas G, Cuppens-Boulahia N, Capitani di Vimercati, eds. *Lecture notes in computer science: data privacy management and autonomous spontaneous security*. Berlin: Springer, 2012:9–24. http://link.springer.com/chapter/10.1007%2F978-3-642-28879-1_2#
- 59 Karakasidis A, Verykios VS. Privacy preserving record linkage using phonetic codes. *Informatics, 2009. BCI '09. Fourth Balkan Conference*; 17–19 September 2009;101, 106.
- 60 Bonomi L, Xiong L, Lu JJ. LinkIT: privacy preserving record linkage and integration via transformations. *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (SIGMOD '13)*. New York, USA: ACM, 2013:1029–32.
- 61 Hall R, Fienberg SE. Privacy-preserving record linkage. *Privacy in Statistical Databases 2010: Lecture Notes in Computer Science*, Volume 6344/2011, 2011:269–83.
- 62 Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. *Info Syst* 2013;38:946–69. ISSN: 0306-4379.
- 63 Christen P. Privacy-preserving data linkage and geocoding: current approaches and research directions. *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference*; December 2006:497, 501.
- 64 Clifton C, Kantarcioglu M, Doan A, *et al.* Privacy-preserving data integration and sharing. *Proceedings of the 9th ACM SIGMOD workshop on research issues in data mining and knowledge discovery (DMKD '04)*. New York, USA: ACM, 2004:19–26.
- 65 Durham E, Xue Y, Kantarcioglu M, *et al.* Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. *Information Fusion* 2012;13:245–59. ISSN: 1566-2535.
- 66 Kuzu M, Kantarcioglu M, Durham EA, *et al.* A practical approach to achieve private medical record linkage in light of public resources. *J Am Med Inform Assoc* 2013;20:285–92.
- 67 Kuzu M, Kantarcioglu M, Inan A, *et al.* Efficient privacy-aware record integration. *EDBT/ICDT*, 2013.
- 68 Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. *Annu Rev Public Health* 2011;32:91–108.
- 69 D'Arcy C, Holman J, Bass AJ, *et al.* Population-based linkage of health records in Western Australia: development of a health services research linked database. *Aust N Z J Public Health* 1999;23:453–9.
- 70 Holman C, D'Arcy J, Bass John A, *et al.* A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Aust Health Rev* 2008;32:766–77.
- 71 Boyd JH, Ferrante AM, O'Keefe CM, *et al.* Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC Health Serv Res* 2012;12:480.
- 72 Blakely T, Woodward A, Salmond C. Anonymous linkage of New Zealand mortality and census data. *Aust N Z J Public Health* 2000;24:92–5.

- 73 Ford D, Jones K, Verplancke J-P, *et al.* The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009;9:157.
- 74 International Health Data Linkage Network (IHDNLN). <http://www.ihdln.org/data-linkage-centres>
- 75 Hertzman CP, Meagher N, McGrail KM. Privacy by Design at Population Data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest. *J Am Med Inform Assoc* 2013;20:25–8.
- 76 Kelman CW, Bass AJ, Holman CDJ, *et al.* Research use of linked health data—a best practice protocol. *Aust N Z J Public Health* 2002;26.3:251–5.
- 77 Pommerening K, Miller M, Schmidtman I, *et al.* Pseudonyms for cancer registries. *Methods Inf Med* 1996;35:112–21.
- 78 Alhaqbani B, Fidge C. Privacy-preserving electronic health record linkage using pseudonym identifiers. *e-Health Networking, Applications and Services, 2008. HealthCom 2008. 10th International Conference*. 7–9 July 2008: 108, 117.
- 79 Bellare K, Curino C, Machanavajihala A, *et al.* WOO: a Scalable and multi-tenant platform for continuous knowledge base synthesis. *VLDB Endowment* 2013;6. <http://db.disi.unitn.eu/pages/VLDBProgram/pdf/industry/p828-rahurkar.pdf>
- 80 Lane J, Heus P, Mulcahy T. Data access in a cyber-world: making use of cyberinfrastructure. *Trans Data Privacy* 2008;1:2–16.