Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Privacy-preserving record linkage on large real world datasets

Sean M. Randall *, Anna M. Ferrante, James H. Boyd, Jacqueline K. Bauer, James B. Semmens

Centre for Population Health Research, Faculty of Health Sciences, Curtin University, Bentley 6102, WA, Australia

ARTICLE INFO

Article history:
Received 26 June 2013
Accepted 4 December 2013
Available online xxxx

Keywords:
Record linkage
Privacy preserving record linkage
Data integration
Bloom filters
Privacy preserving protocols
Population based research

ABSTRACT

Record linkage typically involves the use of dedicated linkage units who are supplied with personally identifying information to determine individuals from within and across datasets. The personally identifying information supplied to linkage units is separated from clinical information prior to release by data custodians. While this substantially reduces the risk of disclosure of sensitive information, some residual risks still exist and remain a concern for some custodians. In this paper we trial a method of record linkage which reduces privacy risk still further on large real world administrative data. The method uses encrypted personal identifying information (bloom filters) in a probability-based linkage framework. The privacy preserving linkage method was tested on ten years of New South Wales (NSW) and Western Australian (WA) hospital admissions data, comprising in total over 26 million records. No difference in linkage quality was found when the results were compared to traditional probabilistic methods using full unencrypted personal identifiers. This presents as a possible means of reducing privacy risks related to record linkage in population level research studies. It is hoped that through adaptations of this method or similar privacy preserving methods, risks related to information disclosure can be reduced so that the benefits of linked research taking place can be fully realised.

1. Introduction

1.1. Administrative data as resource

Administrative health records, containing information on an individual's health and the health services they have received, cover a large proportion of the population and are generally considered to be highly sensitive data. They are used not only for managing an individual health event, but have important uses in informing research, planning and decision making [1]. Current Australian laws provide a number of safeguards to personal privacy including the requirement that the public benefit in using health information for research outweighs with the privacy risks of doing so for the individual [2].

1.2. Record linkage of health information

The process of record linkage is often used to enable researchers to answer questions which require a picture of an individual's health over time. Record linkage is used to identify administrative records belonging to the same person from multiple datasets. In the absence of a unique person identifier, this task is typically carried out using personally identifying information such as name, date of birth and address. As these identifiers can change and/or include errors within or between datasets, probabilistic statistical methods are typically used to ensure high quality links [3]. This linkage process allows researchers to answer questions about the health of individuals over time rather than solely about discrete health events. Research using linked data has resulted in changes to health services delivery and policy [4]. Large scale investment in record linkage infrastructure has occurred in England [5], Scotland [6], Wales [7], Canada [8] and Australia [9] over the last thirty years. Each of these centres has developed linkage expertise which has enabled important research at a population level.

1.3. Record linkage processes and privacy protection

The linkage of different administrative collections across portfolios usually requires the transfer of data to a trusted party or 'linkage unit', which may or may not be external to the data custodians/owners. Various processes and protocols have been developed to protect the privacy of individual and to maintain the security of data.

1.3.1. Separation principle

One method used in many Australian linkage units to reduce privacy risks is to separate data [10]. Under this model, the data is split into personally identifying data (containing information such as name, address and date of birth) and content data (clinical

* Corresponding author.
E-mail addresses: Sean.randall@curtin.edu.au (S.M. Randall), A.Ferrante@curtin.edu.au (A.M. Ferrante), J.Boyd@curtin.edu.au (J.H. Boyd), Jacqui.Bauer@curtin.edu.au (J.K. Bauer), James.Semmens@curtin.edu.au (J.B. Semmens).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.12.003

or service information used for research). The personal identifiers are released to a linkage unit, whose sole role is to determine which records belong to a single person (the release of name-identifying information for research is typically permitted in Australia through exemptions in privacy laws). This is carried out through probabilistic linkage using personal identifiers, typically with a large manual review component. The linkage unit then sends this information back to the custodian, who uses it to supply clinical information to the researcher (see Figure 1).

This method is used by linkage units in WA and NSW to conduct state-based linkages, and has been adopted by the CDL as its best practice national linkage model [9].

1.3.2. Information governance

In addition to the separation principle, linkage units have adopted strong policies and procedures applying to the obtaining, handling, using and disclosing of personally identifying information. This includes an effort to ensure that staff understand their role and responsibilities, that information assets are protected, that policies exist surrounding breaches and disclosure and that information systems place a high priority on security in their design. These policies and procedures have been adopted and developed with input from data custodians.

1.4. Privacy preserving linkage techniques

By separating clinical data from personal identifiers during the linkage process, the risk of revealing sensitive information about individuals is dramatically reduced. Staff conducting the linkage have access only to identifying information, while researchers see only the clinical information relevant for their research questions. Appropriate information governance within linkage units further reduces the risk of information leaks, whether accidentally or maliciously by operators, or as a result of poor business processes.

Nevertheless, some residual risk to privacy remains and, for some data custodians, this is sufficient to prevent the release of personally identifying information to record linkage units. Ideally, such data custodians seek a zero-risk method of providing accurate linked research data without the need to disclose any identifying information to linkage units.

Various techniques known as privacy preserving linkage have been developed to provide lower risk solutions for record linkage. These methods engage in record linkage on encrypted information, and do not require third parties to see personal identifiers. These techniques each differ in their methods, maturity, practicality and suitability for large scale linkages (particularly of low quality data).

Privacy preserving techniques can be classified into two general categories – those that utilise a third party for performing the linkage (three party protocols) and those that do not (two party protocols). Two-party protocols often require a greater amount of necessary communication and computation [11] to compare records, but can be considered more secure as they do not rely on the existence of a trusted third party [12].

In terms of security, privacy preserving techniques generally adopt the same threat model, but differ in the particular privacy techniques used. Nearly all privacy preserving protocols adopt an 'honest-but-curious' threat model [12], whereby parties are expected to try to carry out the protocol correctly, but will also try and find out as much information as they can from any data they receive.

Perhaps the most important criteria in differentiating privacy preserving protocols are around performance features such as linkage quality, scalability and robustness. Privacy preserving protocols range in terms of the comparison techniques applied, from those carrying out an exact match on entire records, to protocols employing string similarity measures on individual fields. Those protocols utilising more fine-grained techniques in determining similarity will typically give higher linkage quality.

Several privacy preserving protocols are being regularly used for routine record linkage. The Australian Institute of Health and Welfare uses the 2nd, 3rd and 5th letters of surname, the 2nd and 3rd letters of forename, the full date of birth and the persons sex to create a 'statistical linkage key' (SLK) which is used to match records [13]. The SLK has been used successfully for a large number of linkages. The Swiss Anonymous Linkage Code [14] creates an identifier from the phonetic codes of first and last name, along with full date of birth and sex. A similar method has been used to conduct linkage in France [15]. Grhanite [16] also uses privacy preserving protocols; like some other systems, it applies a number of pre-processing steps, including phonetic encoding and nickname resolution, before creating their identifier. The process uses these pre-processing steps and fuzzy matching algorithms to produce linkage results that are probabilistic in nature.

In this paper we adopted the bloom filter method for privacy preserving record linkage, developed by Schnell et al. [11]. There were several reasons why we chose this method over other privacy preserving protocols. Firstly the bloom filter approach differs from most other privacy preserving linkage methods in that it is able to measure the similarity between two fields (for instance, between two names) – a method often used in probabilistic record linkage to ensure high quality. Evaluations of privacy preserving string comparison using bloom filters have demonstrated very high quality [11,17], including quality improvements over the SLK and the Swiss anonymous linkage code [18]. Current evaluations have focussed on small data samples, but the method appears adaptable for large-scale record linkage. The method appears robust and well-developed, with a number of papers investigating its security [19] and proposing additions to its method [18,20].

The use of bloom filters was evaluated to determine its suitability for conducting large scale privacy preserving record linkage. Two datasets, comprising in total over 26 million records, were linked using this method, with results compared to the linkage of unencrypted data. A probabilistic linkage framework was adopted to allow large-scale linkage to occur.

2. Method

2.1. Application of bloom filters

To use bloom filters for encrypted record linkage, the personal identifiers need to be encrypted by data custodians. As this process is technically complicated, data custodians would need to be supplied with software that would enable them to encrypt the records. The data custodians involved in the project would agree on a password or pass phrase used to encrypt the data, which would not be shared with the linkage unit. The encrypted data can then be passed to the linkage unit, who can use it to determine which records belong to the same person (see Figure 2).

2.1.1. Creating and comparing bloom filters

An outline of the encryption process presented by Schnell et al. [11] is shown in Fig. 3 along with the method for comparing two encrypted variables which is shown in Fig. 4. Each value (for instance the given name 'SEAN' on one record) is encrypted separately.

A bloom filter begins as an array of a set length, with all array elements set to zero. Firstly, bigrams (overlapping sets of two letters) of the matching variables are created. Padding has been used to give the first and last letters their own bigrams – for instance,

Please cite this article in press as: Randall SM et al. Privacy-preserving record linkage on large real world datasets. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2013.12.003