Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Methodological Review

Publishing data from electronic health records while preserving privacy: A survey of algorithms

Aris Gkoulalas-Divanis [a,*], Grigorios Loukides [b], Jimeng Sun [c]

[a] IBM Research-Ireland, Damastown Industrial Estate, Mulhuddart, Dublin 15, Ireland
[b] School of Computer Science & Informatics, Cardiff University, 5 The Parade, Roath, Cardiff CF24 3AA, UK
[c] IBM Thomas J. Watson Research Center, 17 Skyline Drive, Hawthorne, NY 10532, USA

ARTICLE INFO

Article history:
Received 1 October 2013
Accepted 5 June 2014
Available online xxxx

Keywords:
Privacy
Electronic health records
Anonymization
Algorithms
Survey

ABSTRACT

The dissemination of Electronic Health Records (EHRs) can be highly beneficial for a range of medical studies, spanning from clinical trials to epidemic control studies, but it must be performed in a way that preserves patients' privacy. This is not straightforward, because the disseminated data need to be protected against several privacy threats, while remaining useful for subsequent analysis tasks. In this work, we present a survey of algorithms that have been proposed for publishing structured patient data, in a privacy-preserving way. We review more than 45 algorithms, derive insights on their operation, and highlight their advantages and disadvantages. We also provide a discussion of some promising directions for future research in this area.

1. Introduction

Electronic Medical Record/ Electronic Health Record (EMR/EHR) systems are increasingly adopted to collect and store various types of patient data, which contain information about patients' demographics, diagnosis codes, medication, allergies, and laboratory test results [22,90,63]. For instance, the use of EMR/EHR systems, among office-based physicians, increased from 18% in 2001 to 72% in 2012 and is estimated to exceed 90% by the end of the decade [56].

Data from EMR/EHR systems are increasingly disseminated, for purposes beyond primary care, and this has been shown to be a promising avenue for improving research [63]. This is because it allows data recipients to perform large-scale, low-cost analytic tasks, which require applying statistical tests (e.g., to study correlations between BMI and diabetes), data mining tasks, such as classification (e.g., to predict domestic violence [107]) and clustering (e.g., to control epidemics [117]), or query answering. To facilitate the dissemination and reuse of patient-specific data and help the advancement of research, a number of repositories have been established, such as the Database of Genotype and Phenotype (dbGaP) [89], in the U.S., and the U.K. Biobank [104], in the United Kingdom.

1.1. Motivation

While the dissemination of patient data is greatly beneficial, it must be performed in a way that preserves patients' privacy. Many approaches have been proposed to achieve this, by employing various techniques [43,5], such as cryptography (e.g., [73,55,121,11]) and access control (e.g., [110,71]). However, these approaches are not able to offer patient anonymity (i.e., that patients' private and confidential information will not be disclosed) when data about patients are disseminated [39]. This is because the data need to be disseminated to a wide (and potentially unknown) set of recipients.

Towards preserving anonymity, policies that restrict the sharing of patient-specific medical data are emerging worldwide [91]. For example, in the U.S., the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) [120] outlines two policies for protecting anonymity, namely *Safe Harbor*, and *Expert Determination*. The first of these policies enumerates eighteen direct identifiers that must be removed from data, prior to their dissemination, while, according to the Expert Determination policy, an expert needs to certify that the data to be disseminated pose a low privacy risk before the data can be shared with external parties. Similar policies are in place in countries, such as the U.K. [2] and Canada [3], as well as in the European Union [1]. These policies focus on preventing the privacy threat of *identity disclosure* (also referred to as *re-identification*), which involves the association of an identified individual with their record in the disseminated data.

* Corresponding author.
E-mail address: arisdiva@ie.ibm.com (A. Gkoulalas-Divanis).

http://dx.doi.org/10.1016/j.jbi.2014.06.002
1532-0464/

However, it is important to note that they do not provide *any computational guarantees for thwarting identity disclosure nor aim at preserving the usefulness of disseminated data in analytic tasks*.

To address re-identification, as well as other privacy threats, the computer science and health informatics communities have developed various techniques. Most of these techniques aim at publishing a dataset of patient records, while satisfying certain privacy and data usefulness objectives. Typically, privacy objectives are formulated using privacy models, and enforced by algorithms that transform a given dataset (to facilitate privacy protection) to the minimum necessary extent. The majority of the proposed algorithms are applicable to data containing demographics or diagnosis codes,[1] focus on preventing the threats of *identity*, *attribute*, and/or *membership* disclosure (to be defined in subsequent sections), and operate by transforming the data using *generalization* and/or *suppression* techniques.

1.2. Contributions

In this work, we present a survey of algorithms for publishing patient-specific data in a privacy-preserving way. We begin by discussing the main privacy threats that publishing such data entails, and present the privacy models that have been designed to prevent these threats. Subsequently, for each privacy threat, we provide a survey of algorithms that have been proposed to block it. When selecting the privacy algorithms to be surveyed in the article, we put preference on methods that have appeared in major conferences and journals in the area, as well as are effective in terms of preserving privacy and maintaining good utility. We opted for discussing algorithms that significantly differ from one another, by excluding articles that propose minor algorithmic variations. For the surveyed privacy algorithms we explain the strategies that they employ for: (i) transforming data, (ii) preserving data usefulness, and (iii) searching the space of potential solutions. Based on these strategies, we classify over 45 privacy algorithms. This allows deriving interesting insights on the operation of these algorithms, as well as on their advantages and limitations. In addition, we provide an overview of techniques for preserving privacy that are designed for different settings and types of data, and identify a number of important research directions for future work.

To the best of our knowledge, this is the first survey on algorithms for facilitating the privacy-preserving sharing of structured medical data. However, there are surveys in the computer science literature that do not focus on methods applicable to such data [39], as well as surveys that focus on privacy preservation methods for text data [94], privacy policies [91,93], or system security [36] issues. In addition, we would like to note that the aim of this paper is to provide insights on the tasks and objectives of a wide range of algorithms. Thus, we have omitted the technical details and analysis of specific algorithms and refer the reader to the publications describing the algorithms for them.

1.2.1. Organization

The remainder of this work is organized as follows. Section 2 presents the privacy threats and models that have been proposed for preventing them. Section 3 discusses the two scenarios for privacy-preserving data sharing. Section 4 surveys algorithms for publishing data, in the non-interactive scenario. Section 5 discusses other classes of related techniques. Section 6 presents possible directions for future research, and Section 7 concludes the paper.

[1] These algorithms deal with either relational or transaction (set-valued) attributes. However, following [34,75,76,87], we discuss them in the context of demographic and diagnosis information, which is modeled using relational and transaction attributes, respectively.

2. Privacy threats and models

In this section, we first discuss the major privacy threats that are related to the disclosure of individuals' private and/or sensitive information. Then, we present privacy models that can be used to guard against each of these threats. The importance of discussing privacy models is twofold. First, privacy models can be used to evaluate how safe data are prior to their release. Second, privacy models can be incorporated into algorithms to ensure that the data can be transformed in a way that preserves privacy.

2.1. Privacy threats

Privacy threats relate to three different types of attributes, *direct identifiers*, *quasi-identifiers*, and *sensitive attributes*. Direct identifiers are attributes that can explicitly re-identify individuals, such as name, mailing address, phone number, social security number, other national IDs, and email address. On the other hand, quasi-identifiers are attributes which *in combination* can lead to identity disclosure, such as demographics (e.g., gender, date of birth, and zip code) [109,128] and diagnosis codes [75]. Last, sensitive attributes are those that patients are not willing to be associated with. Examples of these attributes are specific diagnosis codes (e.g., psychiatric diseases, HIV, cancer, etc.) and genomic information. In Table Table 1, we present an example dataset, in which `Name` and `Phone Number` are direct identifiers, `Date of birth`, `Zip Code`, and `Gender` are quasi-identifiers, and `DNA` is a sensitive attribute.

Based on the above-mentioned types of attributes, we can consider the following classes of privacy threats:

- *Identity disclosure (or re-identification)* [112,128]: This is arguably the most notorious threat in publishing medical data. It occurs when an attacker can associate a patient with their record in a published dataset. For example, an attacker may re-identify `Maria` in Table 1, even if the table is published deprived of the direct identifiers (i.e., `Name` and `Phone Number`). This is because `Maria` is the only person in the table who was born on `17.01.1982` and also lives in zip code `55332`.
- *Membership disclosure* [100]: This threat occurs when an attacker can infer with high probability that an individual's record is contained in the published data. For example, consider a dataset which contains information on only HIV-positive patients. The fact that a patient's record is contained in the dataset allows inferring that the patient is HIV-positive, and thus poses a threat to privacy. Note that membership disclosure may occur even when the data are protected from identity disclosure, and that there are several real-world scenarios where protection against membership disclosure is required. Such interesting scenarios were discussed in detail in [100,101].
- *Attribute disclosure (or sensitive information disclosure)* [88]: This threat occurs when an individual is associated with information about their sensitive attributes. This information can be, for example, the individual's value for the sensitive attribute (e.g., the value in *DNA* in Table 1), or a range of values which contain an individual's sensitive value (e.g., if the sensitive attribute is *Hospitalization Cost*, then knowledge that a patient's value in this attribute lies in a narrow range, say $[5400, 5500]$, may be considered as sensitive, as it provides a near accurate estimate of the actual cost incurred, which may be considered to be high, rare, etc.).

There have been several incidents of patient data publishing, where identity disclosure has transpired. For instance, Sweeney [112] first demonstrated the problem in 2002, by linking a claims