Journal of Biomedical Informatics 65 (2017) 76–96

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints

Giorgos Poulis [a,*], Grigorios Loukides [b], Spiros Skiadopoulos [a], Aris Gkoulalas-Divanis [c]

[a] Department of Informatics and Telecommunications, University of the Peloponnese, Greece
[b] Department of Informatics, King's College London, UK
[c] IBM Watson Health, Cambridge, MA, USA

ARTICLE INFO

Article history:
Received 5 February 2016
Revised 22 October 2016
Accepted 1 November 2016
Available online 8 November 2016

Keywords:
Privacy
Demographics
Diagnosis codes
Utility constraints
Generalization
Suppression

ABSTRACT

Publishing data about patients that contain both demographics and diagnosis codes is essential to perform large-scale, low-cost medical studies. However, preserving the privacy and utility of such data is challenging, because it requires: (i) guarding against identity disclosure (re-identification) attacks based on both demographics and diagnosis codes, (ii) ensuring that the anonymized data remain useful in intended analysis tasks, and (iii) minimizing the information loss, incurred by anonymization, to preserve the utility of general analysis tasks that are difficult to determine before data publishing. Existing anonymization approaches are not suitable for being used in this setting, because they cannot satisfy all three requirements. Therefore, in this work, we propose a new approach to deal with this problem. We enforce the requirement (i) by applying (k, k^m)-anonymity, a privacy principle that prevents re-identification from attackers who know the demographics of a patient and up to m of their diagnosis codes, where k and m are tunable parameters. To capture the requirement (ii), we propose the concept of utility constraint for both demographics and diagnosis codes. Utility constraints limit the amount of generalization and are specified by data owners (e.g., the healthcare institution that performs anonymization). We also capture requirement (iii), by employing well-established information loss measures for demographics and for diagnosis codes. To realize our approach, we develop an algorithm that enforces (k, k^m)-anonymity on a dataset containing both demographics and diagnosis codes, in a way that satisfies the specified utility constraints and with minimal information loss, according to the measures. Our experiments with a large dataset containing more than 200,000 electronic health records show the effectiveness and efficiency of our algorithm.

1. Introduction

Healthcare organizations collect increasingly large amounts of data, including clinical trials, Electronic Health Records (EHR), disease registries, and medical imaging databases. In fact, the estimated amount of healthcare data in the world was 0.5 Exabytes ($0.5 \cdot 10^{18}$ bytes) in 2012 and is expected to reach 25 Exabytes by 2020 [66]. Healthcare data are essential for performing large-scale, low-cost analyses [18], which range from Genome-Wide Association Studies (GWAS) to predictive modeling [9,32] and have the potential to improve medical research and practice. For instance, the study in [14] used more than 350,000 records of the Scandinavian Donations and Transfusions database, along with the donors' and the recipients' health records, to answer whether blood transfusions transmit cancer, and it had a substantial impact on public health policies regarding restrictions placed on blood donors [3,13]. Another study [75] used an EHR database of over 300,000 records, to learn meaningful comorbidities, which are associated with different stages of Chronic Obstructive Pulmonary Disease (COPD). This study has the potential to improve COPD prognosis, drug development, and clinical trial design.

While the value of analyzing healthcare data is widely recognized, data sharing remains an obstacle for the majority of healthcare providers [17]. In particular, the privacy-preserving sharing of healthcare data beyond authorized recipients (e.g., researchers or employees of the institution that has collected the data) is challenging [15,16,25]. This is partly because it cannot be facilitated based on access control and encryption-based methods [59,65], or by relying solely on policies (e.g., the HIPAA Privacy Rule [57] in the US, the Anonymization Code [56] in the UK, and the Data Protection Directive [58] in the EU). In fact, a major concern in

* Corresponding author.
E-mail addresses: poulis@uop.gr (G. Poulis), gloukides@acm.org (G. Loukides), spiros@uop.gr (S. Skiadopoulos), gkoulala@us.ibm.com (A. Gkoulalas-Divanis).

http://dx.doi.org/10.1016/j.jbi.2016.11.001
1532-0464/

healthcare data publishing is *identity disclosure* (or *re-identification*), an attack in which patients are linked with their records in the published dataset. Identity disclosure can be performed, even when the published dataset is devoid of direct identifiers (e.g., patient phone numbers), due to the availability of external data sources that can be linked to the published dataset, based on demographics [67] or diagnosis codes [47]. For example, Sweeney estimated that 87% of US citizens can be re-identified based on gender, date of birth, and ZIP code, while Golle [27] estimated this percentage as 63%, using newer, Census 2000 data. In addition, Loukides et al. [44] showed that 96% of 2700 patients, who are involved in an NIH-funded GWAS, are uniquely re-identifiable based on their set of diagnosis codes. In response, various methods have been proposed to prevent identity disclosure when publishing a dataset that contains demographic attributes (e.g., [16,39,64,79]), or diagnosis codes (e.g., [24,31,47,70]).

In this work, we consider the problem of preventing identity disclosure when we need to publish datasets containing both demographics and diagnosis codes, henceforth termed *RT*-datasets. Such datasets are used in many applications [61]. Here, we provide some recent examples:

1. The CMS-HCC risk adjustment model [28] uses demographics and diagnoses of health insurance beneficiaries, to predict the health costs of a US health insurance program, called Medicare Advantage. In particular, beneficiaries' demographics (e.g., gender, aged/disabled status, and whether a beneficiary lives in a certain community or close to an institution) and diagnostic data are used to build and update the risk model. The data are provided from hospital inpatient, hospital outpatient, and physicial risk adjustment data.
2. Various epidemiological [11,26,62] and cancer research [36] studies are based on data containing demographics and diagnosis codes of patients in New South Wales (NSW), Australia. For example, the study of [36] used the data of women over 45 who are associated with certain diagnosis and procedural codes indicating invasive breast cancer. These data were obtained from the NSW Cancer Registry and from several routinely-collected administrative and self-reported health datasets in NSW, and they were analyzed to find out their predictive power in identifying invasive breast cancer cases.
3. The study of [7] uses a dataset containing demographics and ICD-9 diagnosis codes of patients from various US hospitals, to identify groups of patients that are likely to be diagnosed with diabetes, based on their demographics. In particular, it uses multi-label learning algorithms [74], to estimate the risk that each patient has for being diagnosed with diabetes, based on multiple demographics, such as race, gender, and age group.

These applications use data that are devoid of direct identifiers and thus potentially susceptible to identity disclosure. However, their authors recognize the need for algorithms that anonymize both demographics and diagnosis codes, in order to prevent identity disclosure [7] and increase data availability [36]. Also, publishing *RT*-datasets is important to support analysis tasks, including case count studies [46,54], which require accurately counting the number of patients associated with specific demographics and diagnosis codes, predictive modeling, and query answering [61].

However, anonymizing an *RT*-dataset in a utility-preserving way is a very challenging task. This was acknowledged in [54], which is the first work that studied the general problem of anonymizing an *RT*-dataset. As explained in Section 2, our work differs from that of [54] in terms of five main dimensions (anonymization principle, data transformation operation, support of utility constraints, information loss criterion, and anonymization algorithm). Specifically, there are three challenges entailed in the anonymization of an *RT*-dataset in a utility-preserving way. First, identity disclosure cannot be prevented by applying existing methods on demographics and on diagnosis codes separately. This is because an attacker with knowledge of both demographics and diagnosis codes can still re-identify a patient, when the combination of demographics and diagnosis codes of the patient is unique in the anonymized dataset [61]. Specifically, the probability of re-identifying a patient based on such a combination is the reciprocal of the frequency of the combination in the anonymized dataset. Second, data utility must be preserved. This requires constructing an anonymized *RT*-dataset which allows performing: (i) intended analysis tasks with no loss of accuracy and (ii) general analysis tasks, which are difficult to determine before data publishing, with minimum loss of accuracy.

However, the existing methods for anonymizing *RT*-datasets [54,61,68] may substantially reduce the accuracy of intended analysis tasks, or incur excessive information loss, which reduces the accuracy of general analysis tasks. Specifically, the method of [54] does not preserve data *truthfulness*, because it is based on noise addition. That is, it produces synthetic data. Such data are useful for general statistical analysis or mining tasks and can offer strong privacy guarantees [54]. However, the fact that they contain fake information about patients makes them unsuitable for certain applications. For example, they may lead to false alarms in epidemiology [8]. Therefore, our focus is on an anonymization approach that produces truthful data. In addition, the methods of [61,68] do not preserve both aspects of data utility; the output of [61] is of little use in intended tasks and that of [68] incurs substantial information loss, which affects the output of general analysis tasks. To illustrate the challenges of anonymizing an *RT*-dataset, we provide Example 1.

Example 1. Consider the *RT*-dataset D in Fig. 1a. Age, Origin and Gender are demographic attributes, and Disease is a set containing diagnosis codes, whose description is presented in Fig. 2. The dataset in Fig. 1b was produced by applying the method of [78] on the demographic attributes and the method of [45] on Disease. In particular, the latter dataset satisfies 2-anonymity [67] and 2^2-anonymity [50,71], because no record contains a unique combination of demographic values, or a unique combination of two or fewer diagnosis codes (the result of generalizing diagnosis codes is enclosed in () and interpreted as any combination of the codes). However, an attacker who knows the demographics and two diagnosis codes of a patient can still re-identify patients. For example, an attacker who knows that Zoe is a 30-year-old Female from Spain, diagnosed with 493.2 (Chronic obstructive asthma) and 494.1 (Bronchiectasis with acute exacerbation), can associate her with the record 3 of the dataset in Fig. 1b.

Consider also that the *RT*-dataset in Fig. 1a needs to support a study which requires counting the number of patients who are at most 50 and are diagnosed with 494.1. Applying the method of [61] (respectively, of [68]) produces the anonymized dataset in Fig. 1c (respectively, in Fig. 1d). However, these datasets cannot support the study, since the number of records corresponding to patients at most 50 who are diagnosed with 494.1 cannot be accurately determined. This is because the Age values of the records 0 to 4 have been replaced with the range (interval) [19:51] in Fig. 1c, while 494.1 has been generalized together with other diagnosis codes in Fig. 1d.

To address these challenges, our work makes the following specific contributions.

1 **Utility constraints for *RT*-datasets.** We investigate how to model and enforce the requirement of supporting case count studies with no accuracy loss. We propose the concept of utility constraint for *RT*-datasets, building upon previous work on