Journal of Biomedical Informatics 52 (2014) 243–250

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Size matters: How population size influences genotype–phenotype association studies in anonymized data

Raymond Heatherly [a,*], Joshua C. Denny [a,b], Jonathan L. Haines [c], Dan M. Roden [b,d], Bradley A. Malin [a,e]

[a] Department of Biomedical Informatics, School of Medicine, Vanderbilt University, 2525 West End Avenue, Suite 1030, Nashville, TN 37203, USA
[b] Department of Medicine, School of Medicine, Vanderbilt University, 2525 West End Avenue, Suite 1030, Nashville, TN 37203, USA
[c] Department of Epidemiology and Biostatistics, University School of Medicine, Case Western Reserve University, USA
[d] Department of Pharmacology, School of Medicine, Vanderbilt University, 2525 West End Avenue, Suite 1030, Nashville, TN 37203, USA
[e] Department of Electrical Engineering and Computer Science, School of Engineering, Vanderbilt University, 2525 West End Avenue, Suite 1030, Nashville, TN 37203, USA

ARTICLE INFO

Article history:
Received 25 October 2013
Accepted 7 July 2014
Available online 16 July 2014

Keywords:
Privacy
Anonymization
Clinical codes
Data publishing

ABSTRACT

Objective: Electronic medical records (EMRs) data is increasingly incorporated into genome–phenome association studies. Investigators hope to share data, but there are concerns it may be "re-identified" through the exploitation of various features, such as combinations of standardized clinical codes. Formal anonymization algorithms (e.g., k-anonymization) can prevent such violations, but prior studies suggest that the size of the population available for anonymization may influence the utility of the resulting data. We systematically investigate this issue using a large-scale biorepository and EMR system through which we evaluate the ability of researchers to learn from anonymized data for genome–phenome association studies under various conditions.

Methods: We use a k-anonymization strategy to simulate a data protection process (on data sets containing clinical codes) for resources of similar size to those found at nine academic medical institutions within the United States. Following the protection process, we replicate an existing genome–phenome association study and compare the discoveries using the protected data and the original data through the correlation (r^2) of the p-values of association significance.

Results: Our investigation shows that anonymizing an entire dataset with respect to the population from which it is derived yields significantly more utility than small study-specific datasets anonymized unto themselves. When evaluated using the correlation of genome–phenome association strengths on anonymized data versus original data, all nine simulated sites, results from largest-scale anonymizations (population $\sim 100,000$) retained better utility to those on smaller sizes (population $\sim 6000—75,000$). We observed a general trend of increasing r^2 for larger data set sizes: $r^2 = 0.9481$ for small-sized datasets, $r^2 = 0.9493$ for moderately-sized datasets, $r^2 = 0.9934$ for large-sized datasets.

Conclusions: This research implies that regardless of the overall size of an institution's data, there may be significant benefits to anonymization of the entire EMR, even if the institution is planning on releasing only data about a specific cohort of patients.

1. Introduction

Large-scale genotype–phenotype association studies have rapidly increased in prevalence, due to a combination of massively high-throughput technologies [1], lower cost computing platforms, and systems that make information more widely available (e.g., the Database of Genotypes and Phenotypes (dbGaP) [2]). At the same time, it has been shown that data residing in electronic medical records (EMRs) can enable such studies [3–6] finding, for instance, associations with atrioventricular conduction [7], white [8] and red [9] blood cell traits, hypothyroidism [10], and, more recently, the study of pharmacogenetic traits, including clopidogrel-response [11] and warfarin dose [12]. This is further notable because there are indications that learned associations can enable more effective and safe healthcare [13], with early gains in drug dosing [14].

Given the increased reliance upon EMRs for big data research projects, it is important for institutions to work towards data sharing strategies [15–17]. Beyond adhering to policy requirements [18], data sharing can support a wide range of activities [19], including validation of published findings and discovery of novel associations [20]. Despite the opportunities that biomedical data

[*] Corresponding author.
E-mail address: r.heatherly@vanderbilt.edu (R. Heatherly).

http://dx.doi.org/10.1016/j.jbi.2014.07.005
1532-0464/© 2014 Elsevier Inc. All rights reserved.

sharing holds, there are significant concerns over the privacy implications [21,22].

As part of a data protection plan, it is often suggested that biomedical data be disseminated in a manner, such that it is "de-identified" or devoid of explicit identifiers (e.g., personal names) [18,23]. Over the past decade a growing list of investigations have called into question the extent to which de-identification can guard research participants engaged in genomic studies from unsanctioned "re-identification" due to the very act of releasing genomic information itself [24–28]. While we admit that this is an area of concern [29], the likelihood that such attacks will be realized in practice is currently unknown. Thus, in this work, we focus upon linkage risks posed by information that, at the present moment, is more likely to be exploited in re-identification attacks [30].

For years, it has been known that certain common demographics, such as date of birth, gender, and 5-digit ZIP code, could be exploited to discern an individual's identity [31–34]. And, even when demographics are appropriately protected, it may be possible to exploit other features, such as standardized clinical information. This is a concern because it has been illustrated that the set of insurance billing codes (e.g., International Classification of Diseases (ICD)) in a patient's record are often unique [35]. And, while the abstraction of billing codes (e.g., changing of a code representing that a patient suffered from malignant neoplasm of thyroid gland (code 193) to that of neoplasm (codes 140–249)) can drastically reduce the identifiability of patients within a genome–phenome association studies, it can also have a detrimental impact on the underlying data. Ref. [36] proposed a method of clinical code anonymization that yielded data appropriate for validation of reported findings. However, the attack scenario invoked in that work assumed that an adversary has an almost-complete knowledge of the sample population that is being published (e.g., which individuals in a population were included in the study). While possible, the strength of this attacker may not be reasonable, such that instead, in a scenario where the institution publishing the data has a higher level of (but not complete) trust in the system and recipients of data, we consider a modified attacker with a more limited set of knowledge.

An initial examination of the ability to protect clinical data (in regards to re-identifiability) was provided in [37]. In that work, the effects of protection were examined at three naturally-occurring levels within a large academic healthcare system: (1) all patients in the EMR systems, (2) all patients with specimens in a biorepository (a subset of the EMR), and (3) a cohort of patients whose DNA and EMRs were studied to validate certain genotype–phenotype associations (a subset of the biorepository). In these scenarios, the attacker was an individual with knowledge of a patient's visit to the healthcare institution, and their goal was to identify the patient within the published data. It was observed that by protecting a study's cohort with respect to the entire group of patients within the system, the disclosed data could support the discovery of findings with significance that exactly match those of the association observed in the original system.

These findings suggested such a protection method is viable, but the study was limited because it was evaluated in a specific setting. In particular, it was not clear how these findings might translate to other institutions. For instance, at the time of this study, the biobank of the Vanderbilt University Medical Center (VUMC), contained on the order of 110,000 specimens; yet other institutions involved in the Electronic Medical Records and Genomics (eMERGE) network have considerably smaller biorepositories than VUMC which has approximately 110,000 records in its biorepository[1] (e.g., Northwestern University has approximately 15,000 records, and the Mayo Clinic has approximately 20,000). Additionally, other repositories aim for a significant larger population, such as UK Biobank, which plans on over 500,000 participants [38].

Thus, in this paper, we examine the issue that other institutions may face when confronted with the prospect of sharing data – namely, that their overall EMR and biorepository may not be the same size or bias (as regards composition of patients in the biorepository versus in the more general hospital population) as was investigated in [37]. For example, two institutions may be developing a biorepository of a similar size. If, however, Institution A targets their development toward a specific phenotype (e.g., congestive heart failure) while Institution B develops a general-use repository, these biobanks will have different biases (e.g., the rate of appearance for ICD codes representative of CHF will be significantly greater in the former) and potentially even different repository sizes.

To perform this investigation, we conduct a large-scale sensitivity analysis between privacy (that is, "Can an individual patient be re-identified from published data?") and utility ("Is the data usable in various genome–phenome association studies?"). We examine how anonymizing different quantities of electronic medical record data and biorepositories (from small groups of 1000 individuals up to a biorepository of 100,000 individuals and an EMR of over 1,000,000 individuals) affects the results of genome–phenome associations after application of a formal data anonymization algorithm.

The remainder of this paper is structured as follows: in Section 2, we discuss the relevant background to the anonymization approach. In Section 3, we review the anonymization algorithm, describe the experimental process, and detail the measures by which we analyze the algorithm. In Section 4, we highlight the results of the experiments and provide insight into their implications. In Section 5, we provide some intuition into the larger implications of this work and potential future directions of study.

2. Background

2.1. Data privacy and policy

The protection of data derived from EMRs for release happens at multiple levels (e.g., federal law, state law, and institutional policies). Within the United States, the Health Insurance Portability and Accountability Act (HIPAA) provides de-identification specifications at the federal level [39]. These guidelines seek to prevent the unique identification of individuals in published data (i.e., identity disclosure). The HIPAA Privacy Rule offers two alternative approaches to achieve de-identification: (1) Safe Harbor and (2) Expert Determination. When using the Safe Harbor policy, all explicit identifiers (e.g., patient names, Social Security numbers, and medical record numbers) are completely removed and quasi-identifiers (or QIDs) are either removed or abstracted to more general concepts. However, residual information contained within Safe Harbor-compliant data may be exploited by the users, provided they have sufficient background knowledge. For example, as alluded to earlier, it has been shown that even a few visits' worth of ICD-9 codes uniquely identify individuals within an EMR system in [35]. Given such vulnerabilities, it has been suggested that more attention should be paid to the second method for de-identification [40]. In Expert Determination, data is said to be de-identified, when an expert deems that there is "very small" risk that the anticipated recipient of the data could uniquely identify the corresponding individual from which the data was derived. Here, we focus on a method of data protection within the Expert Determination scope.

[1] These values correspond to the size of the biorepositories at the end of the first phase of the eMERGE network in 2011.