Published in final edited form as:

ITAB Corfu Greece (2010). 2010 November ; 2010: 1–6. doi:10.1109/ITAB.2010.5687720.

Privacy-Preserving Publication of Diagnosis Codes for Effective Biomedical Analysis

Grigorios Loukides,

The Department of Biomedical Informatics, Vanderbilt University, USA

Aris Gkoulalas-Divanis, and

IBM research, Zurich, Switzerland

Bradley Malin

The Department of Biomedical Informatics, Vanderbilt University, USA

Grigorios Loukides: grigorios.loukides@vanderbilt.edu; Aris Gkoulalas-Divanis: agd@zurich.ibm.com; Bradley Malin: b.malin@vanderbilt.edu

Abstract

Patient-specific records contained in Electronic Medical Record (EMR) systems are increasingly combined with genomic sequences and deposited into bio-repositories. This allows researchers to perform large-scale, low-cost biomedical studies, such as Genome-Wide Association Studies (GWAS) aimed at identifying associations between genetic factors and complex health-related phenomena, which are an integral facet of personalized medicine. Disseminating this data, however, raises serious privacy concerns because patients' genomic sequences can be linked to their identities through diagnosis codes. This work proposes an approach that guards against this type of data linkage by modifying diagnosis codes in a way that limits the probability of associating a patient's identity to their genomic sequence. Experiments using EMRs from the Vanderbilt University Medical Center verify that our approach generates data that can support up to 29.4% more GWAS than the best-so-far method, while permitting biomedical analysis tasks several orders of magnitude more accurately.

I. Introduction

Electronic Medical Record (EMR) systems are increasingly adopted in many countries [1], [2] and contain large volumes of patient-level data that can be re-used for research purposes, such as to support a range of data analysis tasks [3], [4]. For instance, EMRs are combined with genomic sequences to enable Genome-Wide Association Studies (GWAS). These studies discover genotype-phenotype associations that can improve diagnosis and treatment [5] and facilitate personalized medicine, but require large patient populations to be applied effectively. Thus, the National Institutes of Health (NIH) in the US requires data involved in all NIH-funded GWAS to be deposited into bio-repositories for broad dissemination [6].

To protect patients' right to privacy, the NIH requires de-identifying the deposited data, i.e., removing attributes, such as names, that can reveal patients' identities [7]. However, this is

insufficient to preserve privacy, because patient identities can be linked to genomic sequences through diagnosis codes. For example, more than 96% of 2700 EMRs from a dataset involved in an NIH-funded GWAS were shown to be identifiable based on their diagnosis codes [8]. This poses a serious privacy threat because diagnosis codes exist in EMR systems and hospital discharge summaries, which are publicly available in the US [8], and identified genomic information may be abused [9].

To illustrate this threat, consider that an institution de-identifies and then disseminates the data of Fig. 1(a), which is involved in a GWAS on Bipolar I disorder, a diagnosis that corresponds to the set of ICD-9 codes {296.00, 296.01, 296.02}. In this data, each record corresponds to a distinct patient and contains a set of ICD-9 codes this patient is diagnosed with and their DNA sequence. An attacker with access to an EMR system (containing patients' names and ICD-9 codes) can associate *Tom* with his DNA sequence using the de-identified data (comprised of ICD-9 codes and DNA sequences), because no other patient in this data is diagnosed with the set of codes {296.00, 296.01, 296.02} that *Tom* is diagnosed with.

This threat can be prevented by replacing potentially identifying diagnosis codes with *generalized terms* that are harbored by a “sufficiently” large number of records [10]. For example, the codes 295.00 and 296.00 may be replaced by (295.00, 296.00), a generalized term indicating a diagnosis of *Simple type schizophrenia, unspecified state and/or Bipolar I disorder single manic episode, unspecified degree*. This process, called *generalization*, helps protecting privacy because it can lower the probability of linking a patient's identity to their DNA sequence through diagnosis codes [11].

Generalization should not affect the findings of GWAS or biomedical analysis tasks, but this requirement is not met by existing methods [12], [13] with the exception of [10]. The approach of [10] generalizes diagnosis codes in a way that preserves a set of associations between codes and genomic information that are specified by a researcher or institution. However, [10] over-generalizes the codes not contained in the specified associations, i.e., it replaces them with “too” abstract generalized terms. Unfortunately, increasing the number of the specified associations does not help, because the larger the number of associations, the more difficult it becomes to preserve them. Thus, the data generated by the method of [10] may not permit a large number of biomedical studies beyond those specified through the provided associations to be performed accurately.

In this work, we propose a new method to anonymize diagnosis codes from a standard terminology, including ICD-9 or ICD-10 codes, that effectively supports both GWAS and general biomedical analysis tasks. This is important because it is difficult for researchers to predict all possible uses of data, such as GWAS for different disorders, after it is deposited into a bio-repository.

Our work makes the following major contributions

First, we propose Clustering-Based Anonymizer (CBA), an effective algorithm to generalize data in a way that it links to no less than k patients with respect to potentially identifying sets of diagnosis codes. CBA leverages a powerful, clustering-based heuristic to preserve data

utility. Consider, for example, releasing the de-identified data of Fig. 1(a) to support two GWAS, one on Schizophrenia and another on Bipolar I disorder and assume that no patient associated with at least one ICD code in $\{295.00, \dots, 295.04\}$ or $\{296.00, 296.01, 296.02\}$ should be uniquely re-identified. When applied to the data of Fig. 1(a) with $k = 2$, CBA generates the data of Fig. 1(b). Observe that a patient is now linked to at least 2 DNA sequences using any subset of ICD codes in $\{295.00, \dots, 295.04\}$ or $\{296.00, 296.01,$

$296.02\}$, which effectively limits the re-identification probability to $\frac{1}{2}$. In addition, the distribution of Bipolar I disorder and Schizophrenia is the same as the original data shown in Fig. 1(a) (e.g., two patients have at least one ICD code in $\{296.00, 296.01, 296.02\}$, which implies that they are diagnosed with Bipolar I disorder in both datasets). This allows supporting GWAS on these diseases.

Second, our experimental evaluation using patient records from Vanderbilt's EMR system verifies that CBA is significantly more effective than the method of [10] in terms of generating practically useful anonymized data. Specifically, in all tested cases, CBA constructed anonymized data that supports a larger number of GWAS than that produced by the method of [10], while allowing biomedical studies focusing on clinical case counts to be performed more accurately by several orders of magnitude.

The rest of the paper is organized as follows. Related work is reviewed in Section II. In Section III, we formally define the privacy and utility concepts used by our approach and the problem we consider. The CBA algorithm is presented in Section IV. Finally, Section V reports experimental results and Section VI concludes the paper.

II. Related Work

There are several privacy principles and algorithms for anonymizing relational data, such as patient demographics (see [14] for a survey). However, as shown in [8], [15], [16], these methods are inadequate for anonymizing the type of data we consider in this work without over-generalizing it. This is because the data we consider has rather different semantics. Specifically, each record is associated with a set of diagnosis codes, and different records can harbor a large, variable number of codes.

Anonymizing data in which records are associated with sets of items, such as purchased items, has recently been considered in [12], [13]. However, these methods are inadequate to generate data that supports biomedical analysis tasks, because they perform anonymization in a way that neglects associations between clinical and genomic information [10]. For example, the methods of [12], [13] generalize 296.00 to $(295.00, 296.00)$, a generalized term indicating a diagnosis of *Simple type schizophrenia, unspecified state* and/or *Bipolar I disorder single manic episode, unspecified degree*, which does not allow a researcher to accurately compute the number of patients diagnosed with *Bipolar I disorder* (i.e., records that have at least one ICD code in $\{296.00, 296.01, 296.02\}$).

On the contrary, [10] proposes an algorithm, called Utility-Guided Anonymization of Clinical Profiles (UGACLIP), to generalize sets of potentially identifying diagnosis codes that are extracted from the data or provided by a researcher or institution. UGACLIP was

designed to preserve the specified associations between codes and genomic information. For example, UGACLIP constructs the generalized term (296.00, 296.01) which allows a researcher to be certain that a patient is diagnosed with Bipolar I disorder. However, UGACLIP can over-generalize diagnosis codes that are not contained in the specified associations, as we discuss and experimentally verify later in the paper. This implies that the anonymized data produced by this method may not adequately support a large number of studies beyond those specified through the provided associations.

III. Anonymization Framework

This section presents the framework which forms the basis of our anonymization methodology. Specifically, we describe the type of data, anonymization principles, and ways to control the amount of generalization we consider in this work. Next, we formulate the problem we study.

A. Preliminaries

We consider anonymizing data in which each record, called *transaction*, corresponds to a distinct patient. A transaction T is associated with a set of ICD codes I . A dataset $\mathcal{D}$ is a set of N transactions. A transaction having a set of ICD codes J supports a set of ICD codes I , if $I \subseteq J$. Given a set of ICD codes I in $\mathcal{D}$, we use the support of I in $\mathcal{D}$, denoted with $\text{sup}(I, \mathcal{D})$, to represent the number of transactions in $\mathcal{D}$ that support I . Consider, for example, the first transaction in the data of Fig. 1(a), which supports the set of codes $J = \{296.00, 296.01\}$, since J is a subset of $\{296.00, 296.01, 296.02\}$. The support of $\{296.00, 296.01, 296.02\}$ in the data of Fig. 1(a) is 1, since there is not other transaction that supports this set of codes.

B. Generalization and Suppression of ICD codes

To safely disseminate ICD codes, we map them to generalized terms [10]. A generalized term is denoted by listing its ICD code(s) in brackets¹ and interpreted as any of the non-empty subsets of ICD codes contained in it. We define the concept of ICD code generalization below.

Definition 3.1—An ICD code generalization is a partition $\tilde{\mathcal{I}}$ of the set of ICD codes $\mathcal{I}$ in a dataset $\mathcal{D}$ in which each ICD code i in $\mathcal{I}$ is mapped to a generalized term $\tilde{i}$ in $\tilde{\mathcal{I}}$ that contains i .

Generalization can help prevent patient re-identification, because a generalized term has an equal or greater support than each of the ICD codes it replaces. For instance, the generalized term (296.00, 296.01) in Fig. 1(b) is supported by 2 transactions, whereas 296.00 and 296.01 are supported by 2 and 1 transactions, respectively.

In this work, we generalize ICD codes using the standard ICD-9 code hierarchy [17]. However, our approach can be used to modify other standardized codes provided that there is a taxonomy (e.g., hierarchy or ontology) for them.

¹For clarity, we drop () from generalized terms containing one ICD code.

In addition to generalization, we use *suppression*, an operation that removes ICD codes from the anonymized data. Formally, the suppression of an ICD code i is a mapping of i to a suppressed term $\tilde{i} = ()$. Clearly, suppression offers protection because an attacker, who knows that a patient is diagnosed with one or more suppressed ICD codes, cannot associate this patient with less than the number of the DNA sequences contained in the dataset [8].

C. Controlling Information Loss

Generalization and suppression incur information loss, which we control in two ways to ensure that anonymized data can be used effectively in applications. First, we use the notion of a utility policy [10]. A utility policy is comprised of sets of ICD codes, called *utility constraints*, each of which declares the allowable generalizations for the ICD codes in it. A utility constraint $\{296.00, 296.01\}$, for example, implies that 296.00 can either be released intact or generalized together with 296.01. When generalization is performed as specified by a utility constraint, we say that this constraint is satisfied. A utility policy is satisfied when all utility constraints in it are satisfied, as explained below.

Definition 3.2—A utility policy $\mathcal{U}$ is *satisfied* for an ICD code generalization $\tilde{I}$ if and only if, for each generalized term $\tilde{i}$ in $\tilde{I}$, all ICD codes mapped to $\tilde{i}$ are contained in a single utility constraint in $\mathcal{U}$.

Consider, for example, a utility constraint $\{296.00, 296.01\}$. This constraint is satisfied in the data of Fig. 1(b), since all the ICD codes in the generalized term $(296.00, 296.01)$ are contained in it. By specifying a utility constraint for a disease, we guarantee that the number of patients diagnosed with this disease (i.e., those having at least one ICD code related to this disease) will be the same before and after anonymization when this utility constraint is satisfied.

Observe for example that two patients are diagnosed with 296.00 and/or 296.01 in both datasets of Fig. 1(a) and 1(b), respectively. Thus, an anonymized dataset that satisfies a utility policy is of practical utility for GWAS, since it preserves all the associations between utility constraints and genomic sequences present in the original data.

Second, we introduce Information Loss Metric (*ILM*) to capture the information loss incurred by anonymization. *ILM* is based on the number and semantical closeness of ICD codes contained in a generalized term, as well as this term's support. This metric penalizes generalized terms comprised of many ICD codes that are semantically distant (e.g., *Simple type schizophrenia* and *Hepatitis*) and appear in many transactions.

Definition 3.3—Given an anonymized version $\tilde{\mathcal{D}}$ of $\mathcal{D}$ and a generalized term $\tilde{i}$, the ILM for $\tilde{i}$ is computed as

$$ILM(\tilde{i}) = \frac{|\tilde{i}|}{|\mathcal{I}|} \times w(\tilde{i}) \times \frac{sup(\tilde{i}, \tilde{\mathcal{D}})}{N}$$

where $|\tilde{\mathcal{I}}|$ denotes the number of ICD codes mapped to $\tilde{\mathcal{I}}$, $|\mathcal{I}|$ the number of ICD codes in $\mathcal{D}$, and w is a user-specified weight that captures the semantical closeness of ICD codes in $\tilde{\mathcal{I}}$

[10]. The ILM for $\tilde{\mathcal{D}}$ is computed as $ILM(\tilde{\mathcal{D}}) = \sum_{\tilde{i} \in \tilde{\mathcal{I}}} ILM(\tilde{i}) + \sum_{\tilde{i} \in \tilde{\mathcal{I}} \mid \tilde{i} = () \in \tilde{\mathcal{S}}} S(\tilde{i})$, where S is a function that penalizes suppressed terms.

D. Problem Statement

Our intention is to disseminate patients' data in a way that achieves both privacy and utility. To satisfy privacy, each patient's identity must be associated to no less than k transactions in the anonymized dataset. To satisfy utility, associations between utility constraints and genomic sequences must be preserved and the information loss incurred to anonymize data must be minimal. The problem we consider in this work is formulated as follows.

Problem Statement—Given a dataset $\mathcal{D}$, a utility policy $\mathcal{U}$, a set $\mathcal{P} = \{p_1, \dots, p_r\}$ of potentially identifying sets of ICD codes, and k , construct an anonymized version $\tilde{\mathcal{D}}$ of $\mathcal{D}$ such that: (1) $sup(p, \tilde{\mathcal{D}}) \geq k$, for each $p \in \mathcal{P}$, (2) utility policy is satisfied, and (3) $ILM(\tilde{\mathcal{D}})$ is minimal.

IV. Clustering-Based Anonymizer (CBA)

The problem we attempt to solve is NP-hard (the proof follows from [16]) and may not have a solution even when any amount of information loss is allowed (e.g., when all ICD codes need to be released intact). To deal with the hardness of this problem, we develop Clustering-Based Anonymizer (CBA), an algorithm that selects sets of ICD codes to be anonymized in a way that resembles clustering. Clustering-based algorithms are effective at anonymizing relational data [18], but have not been applied to transactional data.

CBA anonymizes ICD codes using generalization and suppression in a greedy fashion until each potentially identifying set of ICD codes corresponds to at least k patients. To produce anonymizations that allow accuracy in both GWAS and general biomedical analysis tasks, CBA generalizes ICD codes as specified by the utility policy and in a way that incurs the smallest amount of information loss. Furthermore, it applies suppression to the minimum number of ICD codes required and only when generalizing ICD codes according to the utility policy does not suffice to achieve privacy. The pseudocode of CBA is provided in Algorithm 1.

Algorithm 1 Clustering-Based Anonymizer (CBA)

input: Dataset $\mathcal{D}$, utility policy $\mathcal{U}$, $\mathcal{P}$ comprised of potentially identifying sets of codes, and k

output: Anonymized dataset $\tilde{\mathcal{D}}$

1. $\tilde{\mathcal{D}} \leftarrow \mathcal{D}$
2. Populate a priority queue PQ with all sets of codes in $\mathcal{P}$
3. **while** (PQ is not empty)
4. Retrieve the top-most set of codes p from PQ
5. **foreach** ($i_m \in p$)
6. **if** (i_m is a generalized term)

```

7.       $i_m \leftarrow$  the set of ICD codes mapped to  $i_m$ 
8.  if ( $\text{sup}(p, \bar{\mathcal{D}}) \geq k$ )
9.      remove  $p$  from  $PQ$ 
10. else
11.     while ( $\text{sup}(p, \bar{\mathcal{D}}) < k$ )
12.         find a pair  $\{i_m, i_s\}$  such that  $i_m$  is contained in  $p$ ,
             $i_m$  and  $i_s$  are contained in the same utility
            constraint  $u \in \mathcal{U}$  and  $ILM((i_m, i_s))$  is minimal
13.          $\tilde{r} \leftarrow \text{anonymize}(\{i_m, i_s\}, p)$ 
14.         update  $p$  by replacing  $\{i_m, i_s\}$  with  $\tilde{r}$ 
15.         store the mapping of  $\tilde{r}$  with the set of all ICD codes
            contained in it in  $\{i_m, i_s\}$ 
16.     remove  $p$  from  $PQ$ 
17. return  $\bar{\mathcal{D}}$ 

```

In steps 1 and 2, CBA initializes a temporary dataset $\bar{\mathcal{D}}$ to the original dataset $\mathcal{D}$ and a priority queue PQ by inserting each set of potentially identifying ICD codes contained in $\mathcal{P}$. PQ orders its elements with respect to their support in decreasing order. The main operation of CBA in steps 3-16 protects all sets of potentially identifying codes in $\mathcal{P}$ by increasing their support to at least k . More specifically, in step 3, we retrieve the top-most set of codes p from PQ and, in steps 6-7, we update its elements to reflect generalizations that may have occurred in the previous iterations of CBA (we will discuss this issue later on). Then, we remove p from PQ if its support is at least k (steps 8 – 9) or modify the ICD codes in it to achieve privacy (steps 11-16).

Code modification starts by finding a pair $\{i_m, i_s\}$, where i_m and i_s are either ICD codes or generalized terms, such that: (1) i_m is contained in p , (2) both i_m and i_s are contained in the same utility constraint, and (3) generalizing i_m and i_s together incurs the least amount of information loss (step 11). This is performed by scanning all possible pairs $\{i_m, i_s\}$ in a utility constraint, a process that is conceptually similar to greedy agglomerative hierarchical clustering [19] with codes playing the role of points and generalized terms the role of clusters. Notice that condition (1) is needed to increase the support of p , bringing p close to being protected, condition (2) is needed to ensure that the specified utility policy is satisfied (see Definition 3.2), and condition (3) is needed to keep information loss at a minimum.

Subsequently, we modify $\{i_m, i_s\}$ using a function *anonymize* (step 13), which checks whether the utility policy is still satisfied after generalizing $\{i_m, i_s\}$. If this is the case, it is guaranteed that this generalization will preserve the associations between the utility constraint and genomic information, so *anonymize* updates the dataset $\bar{\mathcal{D}}$ by replacing $\{i_m, i_s\}$ with the generalized term (i_m, i_s) . Otherwise, the utility constraint cannot be satisfied, and thus CBA proceeds by suppressing the ICD codes in $\{i_m, i_s\}$ from p in an iterative fashion until the support of p becomes at least k . Note that this ensures that privacy is still preserved. Finally, *anonymize* returns the resulting generalized term, which is assigned to $\tilde{r}$ in step 17.

Then, we update p (step 14) and store the association between the generalized term $\tilde{r}$ and the ICD codes contained in it (step 15). The latter operation is required to enable the update of

generalized items required in step 7. Steps 5-15 are repeated until the support of the set of potentially identifying codes p becomes at least k , in which case p is removed from PQ in step 16. Finally, the dataset $\bar{D}$ is returned in step 17.

CBA manages to preserve data utility better than the UGACLIP [10] algorithm for two reasons. First, it adopts a clustering-based heuristic that considers generalizing every ICD code or generalized term, whereas UGACLIP selects a fixed ICD code or generalized term for anonymization. The latter strategy is known to degrade the quality of clustering, particularly when data is sparse and clusters are arbitrarily shaped [19], as is often the case with the type of data we consider. Second, CBA suppresses no more ICD codes than required to ensure privacy when a utility constraint cannot be satisfied through generalization, while UGACLIP suppresses all the ICD codes in the utility constraint, which removes ICD codes unnecessarily. As confirmed by our experiments, the above strategies adopted by CBA are particularly effective when utility constraints contain a large number of ICD codes, as in the case of non-GWAS related diseases that are useful for general biomedical analysis tasks.

V. Experimental Results

For our experiments, we used two sets of patient records derived from a de-identified version of Vanderbilt's EMR system [20]. The first is referred to as Vanderbilt Native Electrical Conduction dataset ($VNEC$), was constructed for the purposes of an NIH-sponsored GWAS, it contains 2762 transactions and 5830 distinct ICD codes. The second dataset, called VNEC Known Controls ($VNEC_{KC}$), is derived from $VNEC$ by retaining GWAS-related ICD codes, it contains 1335 patient records and 305 distinct ICD codes. Transactions in both datasets have a maximum and average of 25 and 3.1 ICD codes, respectively. $VNEC$ and $VNEC_{KC}$ are expected to be deposited into the dbGaP repository [6] to support GWAS and biomedical analysis tasks. Anonymizing these datasets is challenging, since more than 40% of their records are uniquely identifiable [10].

We evaluated the effectiveness of CBA by comparing it to UGACLIP. We implemented both algorithms in C++ and executed them on an Intel 2.8GHz machine with 4GB of RAM. Following [10], we configured the algorithms to treat all ICD codes a patient was diagnosed with during a visit as potentially identifying and to use a utility policy that includes ICD codes associated with GWAS-related diseases.

We designed two sets of experiments. The first set evaluates the algorithms in terms of their ability to support intended GWAS, while the second one evaluates the effectiveness of the algorithms in terms of retaining data utility for general biomedical analysis.

A. Effectiveness for supporting GWAS

We examined which of the utility constraints related to GWAS were satisfied in both of the anonymized datasets by setting k to 5, a commonly used value [21], and to 10, a value that trades off some utility for stronger protection. As can be seen in the result shown for $k = 5$ in Table I, the anonymizations generated by CBA satisfied 22.2% and 5.5% more utility constraints than those created by UGACLIP for $VNEC$ and $VNEC_{KC}$, respectively. The

result for $k = 10$, illustrated in Table II, is qualitatively similar to that of Table I with CBA outperforming UGACLIP by a margin of 29.4% for both datasets. This is expected because selecting a fixed ICD code for generalization, as UGACLIP does, incurs a larger amount of information loss when k is large [18]. Thus, CBA is more effective in anonymizing data guided by a utility policy.

B. Effectiveness for supporting general biomedical analysis

We examined the effectiveness of CBA in generating anonymizations that help general biomedical analysis by using 3 information loss measures: (i) *ILM*, (ii) Normalized Certainty Penalty (*NCP*) [18], which is expressed as the weighted average of the information loss of all generalized terms, each of which is penalized based on the number of items of the original data it replaces, and (iii) Average Relative Error (*ARE*) [18], which captures the accuracy of answering a workload of queries on anonymized data. *ARE* reflects the average number of transactions that are retrieved incorrectly as part of query answers and is computed as the mean error of answering a query workload.

We first evaluated information loss using *ILM* for various k values in [2, 25]. Figs. 2 and 3 illustrate the results with respect to *ILM* for *VNEC* and *VNEC_{kc}*, respectively. It can be seen that CBA incurs an amount of information loss that is orders of magnitude lower than UGACLIP across all k values for both datasets. These results, together with that of Fig. 4, which shows the *NCP* scores for *VNEC* (the result for *VNEC_{kc}* is not shown due to space constraints), verify that the clustering-based search strategy employed by CBA is much more powerful than the heuristic used in UGACLIP.

Finally, we investigated whether CBA can produce anonymized data that can effectively support tasks focusing on clinical case counts. We assumed that data recipients, such as researchers accessing a bio-repository, issue queries to learn the number of transactions having a set of ICD codes supported by at least 5% of all the transactions of a dataset. Answering such queries is needed in various biomedical data analysis applications, including association rule mining [3] and classification [22]. However, anonymized data may not allow such queries to be answered accurately because a generalized term can be interpreted as any of the nonempty subsets of ICD codes it contains. Figs. 5 and 6 illustrate the *ARE* scores for *VNEC* and *VNEC_{kc}* respectively. Observe that CBA significantly outperformed UGACLIP, as it generated anonymizations that permit at least 2.5 times more accurate querying answering.

VI. Conclusions and Future Work

Patient-level data, derived by combining EMRs with genomic information needs to be shared in a way that allows effective analysis without revealing patients' identities. In this work, we proposed a novel approach that achieves both of these goals by employing a utility-guided anonymization methodology together with a clustering-based algorithm. Our approach generates anonymizations that support both GWAS and other biomedical analysis tasks significantly better than the current state-of-the-art method, while achieving the same level of privacy.

This work opens up two main directions for future research. First, we intend to automate the selection of a suitable privacy protection level k , which existing methods leave to researchers or institutions, and the construction of a utility policy. Second, we aim to extend our approach to deal with additional patient features that are disseminated together with diagnosis codes and may be exploited for re-identification.

References

1. Virtanen T. The finnish national ehealth archive and the new research possibilities. *Nursing Informatics*. 2009:688–691.
2. Williams MH, Venters G, Venters G, Marwick DH. Developing a regional healthcare information network. *IEEE Transactions on Information Technology in Biomedicine*. 2001; 5(2):177–180. [PubMed: 11420996]
3. Ordonez C. Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine*. 2006; 10(2):334–343. [PubMed: 16617622]
4. Safran C, Bloomrosen M, Hammond W, Labkoff S, Markel-Fox S, Tang P, Detmer D. Toward a national framework for the secondary use of health data. *Journal of the American Medical Informatics Association*. 2007; 14:1–9. [PubMed: 17077452]
5. Gurwitz D, Lunshof J, Altman R. A call for the creation of personalized medicine databases. *Nature Reviews Drug Discovery*. 2006; 5:23–26.
6. Mailman M, Feolo M, Jin Y, Kimura M, Tryka K, B R, et al. The ncbi dbgap database of genotypes and phenotypes. *Nature Genetics*. 2007; 39:1181–1186. [PubMed: 17898773]
7. Health insurance portability and accountability act of 1996 united states public law.
8. Loukides G, Denny J, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association*. 2010; 17:322–327. [PubMed: 20442151]
9. Rothstein M, Epps P. Ethical and legal implications of pharmacogenomics. *Nature Reviews Genetics*. 2001; 2:228–231.
10. Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. *Proceedings of the National Academy of Sciences USA*. 2010; 107:7898–7903.
11. Samarati P. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*. 2001; 13(9):1010–1027.
12. Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set-valued data. *Proceedings of the VLDB Endowment*. 2008; 1(1):115–125.
13. He Y, Naughton JF. Anonymization of set-valued data via top-down, local generalization. *Proceedings of the VLDB Endowment*. 2009; 2(1):934–945.
14. Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: A survey on recent developments. *ACM Computing Surveys*. Dec.2010 42
15. Aggarwal CC. On k-anonymity and the curse of dimensionality. *VLDB '05*. 2005:901–909.
16. Xu Y, Wang K, Fu AWC, Yu PS. Anonymizing transaction databases for publication. *ACM SIGKDD*. 2008:767–775.
17. C. for Disease Control and Prevention. International classification of diseases, ninth revision, clinical modification (icd-9-cm).
18. Xu J, Wang W, Pei J, Wang X, Shi B, Fu AWC. Utility-based anonymization using local recoding. *ACM SIGKDD*. 2006:785–790.
19. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. *ACM Computing Surveys (CSUR)*. 1999; 31:264–323.
20. Roden D, Pulley J, Basford M, Bernard G, Clayton E, Balser J, Masys D. Development of a large scale de-identified dna biobank to enable personalized medicine. *Clinical Pharmacology and Therapeutics*. 2008; 84(3):362–369. [PubMed: 18500243]

21. Emam KE, Dankar FK. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*. 2008; 15(5):627–637. [PubMed: 18579830]
22. Karaolis M, Moutiris J, Hadjipanayi D, Pattichis C. Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Transactions on Information Technology in Biomedicine*. 2010; 14(3):559–566. [PubMed: 20071264]

<i>Patient</i>	Diagnosis Codes	DNA sequence
<i>Tom</i>	296.00 296.01 296.02	CA...T
<i>Mary</i>	295.00 295.01 295.02 295.03 295.04	TA...C
<i>Bob</i>	295.01 295.02	CT...A
<i>Greg</i>	296.00 296.02 295.02 295.03	GC...C
<i>Anne</i>	295.00 295.01 295.02 295.03	CG....A

(a) Original dataset

Diagnosis Codes	DNA sequence
(296.00, 296.01) 296.02	CA...T
295.00 (295.01, 295.04) 295.02 295.03	TA...C
(295.01, 295.04) 295.02	CT...A
(296.00, 296.01) 296.02 295.02 295.03	GC...C
295.00 (295.01, 295.04) 295.02 295.03	CG....A

(b) Anonymized dataset

Fig. 1. Original and anonymized dataset

Fig. 2. ILM vs. k (VNEC)

Fig. 3. *ILM* vs. *k* (*VNEC_{KC}*)

Fig. 4. *NCP* vs. *k* (*VNEC*)

Fig. 5. ARE vs. k (VNEC)

Fig. 6. ARE vs. k ($VNEC_{KC}$)

Table 1
Satisfied utility constraints for $k = 5$ (✓ denotes that a utility constraint is satisfied)

Disease	VNEC			VNEC _{KC}			Disease	VNEC			VNEC _{KC}		
	CBA	UGACLIP	CBA	UGACLIP	CBA	UGACLIP		CBA	UGACLIP	CBA	UGACLIP	CBA	UGACLIP
Asthma	✓	✓	✓	✓	✓	✓	Lung cancer	✓	✓	✓	✓	✓	✓
Attention deficit with hyperactivity	✓		✓				Pancreatic cancer	✓	✓	✓	✓	✓	✓
Bipolar I disorder		✓			✓		Platelet phenotypes	✓		✓		✓	✓
Bladder cancer	✓						Pre-term birth	✓	✓	✓	✓	✓	✓
Breast cancer	✓	✓	✓	✓	✓	✓	Prostate cancer	✓	✓	✓	✓	✓	✓
Coronary disease		✓					Psoriasis	✓		✓		✓	✓
Dental caries	✓	✓	✓	✓	✓	✓	Renal cancer	✓	✓	✓	✓	✓	✓
Diabetes mellitus type-1		✓			✓	✓	Schizophrenia	✓		✓		✓	✓
Diabetes mellitus type-2		✓			✓	✓	Sickle-cell disease	✓		✓		✓	✓

NIH-PA Author Manuscript

NIH-PA Author Manuscript