# ANONYMIZATION OF LONGITUDINAL ELECTRONIC MEDICAL RECORDS FOR CLINICAL RESEARCH

By

Acar Tamersoy

Thesis
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE
in
Biomedical Informatics

August, 2011
Nashville, Tennessee

Approved:
Bradley Malin, Ph.D., Chair
Joshua C. Denny, M.D., M.S.
Thomas A. Lasko, M.D., Ph.D.

# ACKNOWLEDGMENTS

First, I would like to thank my advisor, Dr. Bradley Malin, for his continuous support, guidance and motivation throughout this work. I am also grateful to the other members of my committee, Dr. Thomas Lasko and Dr. Joshua Denny, for their valuable comments and feedback about this thesis. This research was funded by NIH grants R01LM009989 and U01HG004603.

I am deeply thankful to Dr. Grigorios Loukides, Dr. Mehmet Ercan Nergiz, Dr. Yucel Saygin, and Dr. Aris Gkoulalas-Divanis for their constant advice and guidance. I am looking forward to our continuing collaboration in the future.

Finally, I would like to thank my family for their never-ending encouragement. This work would not have been possible without their support.

# TABLE OF CONTENTS

Page

# LIST OF TABLES

Table Page

# LIST OF FIGURES

Figure Page

# CHAPTER I

# INTRODUCTION

Advances in health information technology have facilitated the collection of detailed, patient-level clinical data to enable efficiency, effectiveness, and safety in healthcare operations [1]. Such data are often stored in electronic medical record (EMR) systems [2, 3] and are increasingly repurposed to support clinical research (e.g., [4, 5, 6, 7]). Recently, EMRs have been combined with biorepositories to enable genome-wide association studies (GWAS) with clinical phenomena in the hopes of tailoring healthcare to genetic variants [8]. To demonstrate feasibility, EMR-based GWAS have focused on static phenotypes; i.e., where a patient is designated as disease positive or negative (e.g., [9, 10, 11]). As these studies mature, they will support personalized clinical decision support tools [12] and will require either repeated data (i.e., data with replicated diagnosis information) for improved diagnostic certainty [13] or longitudinal data (i.e., repeated data with temporal information) to understand how treatment influences a phenotype over time [14, 15].

Meanwhile, there are challenges to conducting GWAS on a scale necessary to institute changes in healthcare. First, to generate appropriate statistical power, scientists may require access to populations larger than those available in local EMR systems [16]. Second, the cost of a GWAS - incurred in the setup and application of software to process medical records as well as in genome sequencing - is non-trivial [17]. Thus, it can be difficult for scientists to generate novel, or validate published, associations. To mitigate this problem, the U.S. National Institutes of Health (NIH) encourages investigators to share data from NIH-supported GWAS [18] into the Database of Genotypes and Phenotypes (dbGaP) [19].

This, however, may lead to privacy breaches if patients' clinical or genomic information is associated with their identities. As a first line of defense against this threat, the NIH recommends investigators *de-identify* data by removing an enumerated list of attributes that could identify patients (e.g., personal names and residential addresses) prior to dbGaP submission [20]. However, a patients' DNA may still be *re-identified* via residual demographics [21] and clinical information (e.g., standardized International Classification of Diseases (ICD) codes) [22] as we demonstrate in later chapters.

Methods to mitigate re-identification via demographic and clinical features [23, 24] have been proposed, but they are not applicable to the data considered in this thesis. These methods assume the clinical profile is devoid of temporal or replicated diagnosis information. Consequently, these methods produce data that are unlikely to permit meaningful clinical investigations.

This thesis addresses a number of important questions regarding the anonymization of repeated and longitudinal data. Specifically, our work makes the following specific contributions:

- We propose an algorithm to guard against re-identification attacks on repeated data. We realize our approach in an automated algorithm which attempts to maximize the number of repeated diagnosis codes that can be safely released. We illustrate the effectiveness of our approach in retaining repeated comorbidities using a patient cohort from the Vanderbilt University Medical Center (VUMC) EMR system. The content of this chapter, in edited form, was published in the proceedings of the 2010 American Medical Informatics Association (AMIA) Annual Symposium as a research paper (see [25]).

- We propose a framework to guard against re-identification attacks on longitudinal data. The framework transforms each longitudinal patient record so that it is indistinguishable from a certain set of other patients' records with respect to potentially identifying information. This is achieved by iteratively clustering records and applying *generalization*, which replaces ICD codes and age values with more general values, and *suppression*, which removes ICD codes and age values. We evaluate our approach with several cohorts of patient records from the VUMC EMR system. Our results demonstrate that the anonymized data derived by our framework allow many studies focusing on clinical case counts to be performed accurately. The content of this chapter is currently under peer-review.

The remainder of the thesis is organized as follows. In Chapter II, we review related research on anonymization and its application to biomedical data. In Chapter III, we present our anonymization methodology for repeated data. In Chapter IV, we focus on more complex longitudinal data and propose a framework to formally anonymize such data. In Chapter V, we discuss the extensions of our methods and highlight the limitations of this study. Finally, Chapter VI concludes the thesis.

# CHAPTER II

# RELATED RESEARCH

Re-identification concerns for clinical data via seemingly innocuous attributes were first raised in [26]. Specifically, it was shown that patients could be uniquely re-identified by linking publicly available voter registration lists to hospital discharge summaries via demographics, such as date of birth, gender, and 5-digit residential zip code. The re-identification phenomenon for clinical data has since attracted interest in domains beyond healthcare, and numerous techniques to guard against attacks have emerged (see [27, 28] for surveys). In this chapter, we survey research related to privacy-preserving data publishing, with a focus on biomedical data. We note that the re-identification problem is not addressed by access control and encryption-based methods [29, 30, 31] because regulations such as the HIPAA Privacy rule permit de-identified data to be publicly shared beyond a small number of authorized recipients.

## II.1 Relational Data

We first discuss methods for preventing re-identification when sharing relational data, such as demographics, in which records have a fixed number of attributes and one value per attribute.

The first category of protection methods transforms attribute values so that they no longer correspond to real individuals. Popular approaches in this category are noise addition, data swapping, and synthetic data generation (see [32, 33, 34] for surveys). While such methods generate data that preserve aggregate statistics (e.g., the average age), they do not guarantee data that can be analyzed at the record level. This is a significant limitation that hampers the ability to use these data in various biomedical studies, including epidemiological studies [35] and GWAS [23]. As an example, consider a pharmaceutical researcher mining published patient-level clinical data to discover previously unknown side effects of a drug. The results of such an analysis are of no use if the data contain noisy or randomized patient records [36].

In contrast, methods based on generalization and/or suppression (e.g., [37, 38, 39]) preserve the truthfulness of the original data at the record level. Many of these methods are based on a privacy principle, called $k$-anonymity [26, 37], which states that each record of the published data must be equivalent to at least $k-1$ other records with respect to *quasi-identifiers* (QI) (i.e., attributes that can be linked with external information for re-identification purposes) [40]. To minimize the information loss incurred by anonymization, these methods employ various search strategies, including binary search [37, 38], partitioning [39], clustering [41, 42], and evolutionary search [43]. Furthermore, there exist methods that have been successfully applied by the biomedical community [38, 44].

*k*-Anonymity is a specific realization of a more general data privacy model known as *k*-map [26, 45]. The latter principle enhances data utility by relaxing the *k*-anonymity requirement. Specifically, *k*-map ensures that each record in the published data can be associated with no less than *k* records in the population with respect to a QI.

## II.2 Transactional Data

Next, we turn our attention to approaches that deal with more complex data. Specifically, we consider transactional data, in which records have a large and variable number of values per attribute (e.g., the set of diagnosis codes assigned to a patient during a hospital visit). Transactional data can also facilitate re-identification in the biomedical domain. For instance, de-identified clinical records can be linked to patients based on combinations of diagnosis codes that are additionally contained in publicly available hospital discharge summaries and EMR systems from which the records have been derived [22]. As it was shown in [22], more than 96% of 2700 patient records, collected in the context of a GWAS, are susceptible to re-identification based on diagnosis codes.

From the protection perspective, there are several approaches that have been developed to anonymize transactional data. Notably, Terrovitis et al. [46] proposed the $k^m$-anonymity principle, along with several heuristic algorithms, to prevent attackers from linking an individual to less than *k* records. This model assumes that the adversary knows at most *m* values of any transaction. To anonymize patient records in transactional form, Loukides et al. [23] introduced a privacy principle to ensure that sets of potentially identifying diagnosis codes are protected from re-identification, while remaining useful for GWAS validations. To enforce this principle, they proposed an algorithm that employs generalization and suppression to group semantically close diagnosis codes together in a way that enhances data utility [23, 24].

The work in this thesis differs from the aforementioned research along three principal dimensions. First, we prevent re-identification in repeated and longitudinal data publishing. Second, contrary to the approaches of [23, 24] which employ *k*-anonymity as their privacy principle, our repeated data anonymization method is more general and realizes a variant of *k*-anonymity to enhance data utility. Third, the approaches of [23, 24] group diagnosis codes together whereas our longitudinal data anonymization framework is based on grouping of records, an approach that has been shown to be highly effective in retaining data utility because of the direct recognition of records being anonymized [41, 42, 47].

### II.3 Spatiotemporal Data

Spatiotemporal data are related to the longitudinal data anonymization problem. They are time and location dependent, and these unique characteristics make them challenging to protect against re-identification. Such data are typically produced as a result of queries issued by mobile subscribers to location-based service providers, who, in turn, supply information services based on specific physical locations.

The principle of *k*-anonymity has been extended to anonymize spatiotemporal data. Abul et al. [48] proposed a technique to group at least *k* objects that correspond to different subscribers and appear within a certain radius of the path of every object in the same time period. In addition to generalization and suppression, [48] considered adding noise to the original paths so that objects appear at the same time and spatial trajectory volume. Assuming that the locations of subscribers constitute sensitive information, Terrovitis et al. [49] proposed a suppression-based methodology to prevent attackers from inferring these locations. Finally, Nergiz et al. [47] proposed an approach that employs *k*-anonymity, enforced using generalization together with reconstruction (i.e., randomly sampling specific records from the area covered by the anonymized data) for improved protection. Our heuristics for anonymizing longitudinal data are inspired from [47], however, we employ both generalization and suppression to further enhance data utility, and we do not use reconstruction to preserve data truthfulness.

The aforementioned approaches are developed for anonymizing spatiotemporal data and cannot be applied to longitudinal data due to different semantics. Specifically, the data considered in this thesis record patients' diagnoses and not their locations. Consequently, the objective of our approach is to prevent re-identification based on patients' diagnosis and time information, not to hide their locations.

# CHAPTER III

## ANONYMIZATION OF REPEATED DIAGNOSIS CODES DERIVED FROM ELECTRONIC MEDICAL RECORDS

In this chapter, we present *Greedy Code Censoring* (GCCens) which is the first approach to formally anonymize EMR-derived repeated data. As we noted in Chapter I, such data consist of replicated diagnoses with no temporal information and, thus, are a special case of more complex longitudinal data. This chapter, therefore, serves as a pilot study for anonymizing longitudinal data. More specifically, in this chapter, we first demonstrate the privacy problem, and present the notation and the materials. We then formalize the risk measure, the GCCens algorithm, and the data utility measure. We conclude the chapter with an experimental evaluation of the proposed approach using a patient cohort derived from the EMR system of the VUMC.

| *S* | ICD | DNA |
|---|---|---|
| *1* | 250 | CT...A |
| *2* | 272, 272, 724 | AC..T |
| *3* | 250, 250, 272 | GC..A |

**(a)**

| *P* | ID | ICD |
|---|---|---|
| 1 | Dan | 250 |
| 2 | Bella | 250, 250, 272 |
| 3 | John | 250, 250, 272, 272 |
| 4 | Ada | 401, 401, 401, 401 |
| 5 | Tom | 272, 272, 724 |
| 6 | Alan | 250 |
| 7 | Eric | 272, 724 |

**(b)**

| *Y* | ICD | DNA |
|---|---|---|
| *1* | 250 | CT...A |
| *2* | 272, 724 | AC..T |
| *3* | 250, 272 | GC..A |

**(c)**

| CUL |
|---|
| 0 |
| 0.33 |
| 0.33 |

**(d)**

*Figure 1:* A depiction of the repeated data privacy problem. (a) and (b) depict repeated data and identified EMR, respectively. A 2-map based on the proposed approach and information loss incurred by this protection are depicted in (c) and (d), respectively.

### III.1 Motivating Example

As an example of the problem studied in this chapter, consider the repeated data in Figure 1a. Each record corresponds to a fictional de-identified patient and is comprised of ICD codes and a DNA sequence. The first record, for instance, denotes that a patient was diagnosed with *diabetes mellitus* (code 250) and has the DNA sequence 'CT...A'. The clinical and genomic data are derived from an EMR system and a research project beyond primary care (i.e., they are not contained in the EMR system), respectively. Publishing the data of

Figure 1a could allow a hospital employee with EMR access to associate *Tom* with his DNA sequence. This is because the identified record, shown in Figure 1b, can only be linked to the second record in Figure 1a based on the combination of ICD codes '272, 272, 724'.

## III.2 A Formal Model of the System

Before proceeding, we formalize the privacy problem considered in this chapter. Let $U = \{d_1, ..., d_h\}$ be the set of distinct ICD codes stored in an EMR system. The dataset $P = \{p_1, ..., p_n\}$ represents the medical records of the patient population. Each record $p_i$ is of the form $< ID_i, D_i >$, where $ID_i$ is an identifier associated with a patient and $D_i$ is a set of ICD codes for the patient (which are not necessarily distinct) derived from $U$. Figure 1b depicts a population that is comprised of seven records. For instance, $p_5$ has $ID_5$ = Tom and $D_5$ = $\{272, 272, 724\}$.

A second dataset $S = \{s_1, ..., s_m\}$ represents a sample of patient records to be shared. Each record $s_j$ is of the form $< D_j, DNA_j >$ and corresponds to a patient whose record is in the population. $D_j$ is a set of ICD codes derived from $U$ and $DNA_j$ represents genomic sequence data. For instance, $s_1$ in Figure 1a has $DNA_1 = \{CT...A\}$ and $D_1 = \{250\}$ which was derived from $p_6$.

The re-identification attack we consider assumes an attacker knows the identifying information and ICD codes about a patient whose record is in the sample. This could occur through various routes. First, a data recipient may be an employee of the institution from which the data were derived, with access to the EMR system. The recipient, alternatively, may have knowledge about a neighbor or coworker. Or, in certain cases, a recipient may use public information; e.g., she may link de-identified hospital discharge summaries with identified resources, such as voter registration lists [21, 26].

## III.3 Materials

For this study, we worked with the de-identified version of StarChart, the EMR system of the VUMC [50]. We constructed $P$ using a set of $301,423$ patients' records that contain at least one of the following ICD codes: *250* (diabetes mellitus), *272* (disorders of lipoid metabolism), *401* (essential hypertension), and *724* (other and unspecified disorders of back). We selected these ICD codes because they appear frequently in the EMR system, they are interesting phenotypes, and they are critical covariates for a wide range of other clinical phenotypes [11].

The research sample $S$ contains $2,676$ patient records and was extracted for the purposes of a GWAS on native electrical conduction within the ventricles of the heart [11]. The sample represents a 'heart healthy' group with no prior heart disease, no heart conduction abnormalities, no electrolyte abnormalities, and no use of medications that can interfere with conduction.

A record in $S$, on average, consists of 3.5, 2.3, 4.4, and 2 repeats of the ICD codes 250, 272, 401, and 724, respectively. A record in $P$, on average, consists of 2.2, 1.3, 2.5, and 0.9 repeats of the same ICD codes. It was shown in previous research [22] that this cohort is appropriate for studying privacy threats in samples derived from the VUMC EMR system.

### III.4 Risk Measure

We measure the level of privacy protection afforded to the sample using the distinguishability measure [22]. This measure is applied to determine how many records are susceptible to re-identification based on shared ICD codes. Specifically, given $s_j$, distinguishability (which we refer to as a function *dis*) is equal to the number of records in $P$ that contain all ICD codes in $D_j$. A patient is said to be *uniquely identifiable* if her record has a distinguishability of 1. Distinguishability is the inverse of the probability of re-identifying a patient. For example, in Figure 1a, $dis(272,724) = 2$ because two records (i.e., $p_5$ and $p_7$) in $P$ contain all of these ICD codes. Note the probability of correctly identifying one of these records is $1/2$.

### III.5 Censoring Algorithm

Greedy heuristics are commonly employed to anonymize data due to their ability to retain both privacy and utility [51]. Along these lines, GCCens is designed to limit the number of ICD codes that are released in patient records in a greedy manner. A notable strength of GCCens is that it significantly enhances data utility by employing $k$-map as its privacy principle. In our setting, $S$ satisfies $k$-map when, for each $D_j$ in $S$, $dis(D_j) \geq k$. This ensures that each record in $S$ can be associated with no less than $k$ records in $P$ based on the released ICD codes, and implies that the probability of performing a re-identification attack is no greater than $1/k$ per disclosed record.

The $k$-map model tends to offer 'high' data utility, but assumes no knowledge of whether an individual's record is contained in $S$. However, such knowledge is difficult (if not impossible) to be acquired by an attacker in the context of the data sharing we consider. This is because, typically, a random fraction of

EMRs with identical ICD codes are associated with DNA information and released.

The GCCens algorithm accepts the following inputs: a sample $S$, a population $P$, a privacy parameter $k$ and a set of censoring thresholds $C = \{c_1, ..., c_{|U|}\}$. The algorithm outputs $Y$, a version of $S$ that is $k$-mapped to $P$. The parameter $k$ expresses the minimum allowable number of records in $P$ that can be mapped to a record of $Y$ based on ICD codes, while $C$ is a set of thresholds (called *caps*), each of which corresponds to an ICD code in $U$ (i.e., $c_m$ is the cap value of the ICD code $d_m$) and expresses the maximum allowable number of times a particular ICD code can appear in a record of $Y$. The caps, in effect, act as an initial acceptable censor for the distribution of repeat counts. In this work, we follow the standard assumptions [26, 52] in that we assume $k$ and $C$ are specified by data owners according to their expectations about an attacker's knowledge. We also note that it is possible to specify $C$ automatically by scanning $S$ and recording the maximum number of occurrences of each distinct ICD code in all records.

**Algorithm 1** GCCens($P$, $S$, $k$, $C$)

```
Require: Population $P$, sample $S$, privacy parameter $k$, cap set $C$
Return: $Y$: $k$-mapped version of $S$
 1: $Y \leftarrow preprocess(S)$ such that no ICD code appears more than its cap
 2: while there exists $D_j \in Y$ such that $dis(D_j) < k$ do
 3:     $R \leftarrow \{r_1, ..., r_{|U|}\}$          ▷ $r_e$ is a set associated with the ICD code $d_e$
 4:     for each ICD code $d_m$ do
 5:         for each record $y_n \in Y$ do
 6:             if $d_m$ appears $c_m$ times in $y_n$ then
 7:                 $r_m \leftarrow r_m \cup y_n$
 8:             end if
 9:         end for
10:     end for
11:     $o \leftarrow \text{argmin}_{f \in \{1...|U|\}}\{|r_f|\}$
12:     for each record $y_t \in r_o$ do
13:         remove $d_o$ from $y_t$
14:     end for
15:     $c_o \leftarrow c_o - 1$
16: end while
17: return $Y$
```

The pseudocode of GCCens is illustrated in Algorithm 1. In step 1, the algorithm invokes a helper function called $preprocess()$, which iteratively censors ICD codes (i.e., removes one of their instances) from $S$ until there is no ICD code that appears more than its cap in $S$. The result of $preprocess()$ is assigned to a sample $Y$. Then, in steps 2-16, GCCens iterates over $Y$ until $k$-map is satisfied. More specifically, in step 3, GCCens generates a set of sets, each of which corresponds to an ICD code in $U$. Then, in steps 4-10, GCCens computes the number of ICD code instances that need to be censored for each distinct ICD code $d_m$. This is achieved by iterating over all records in $Y$ (step 5-9), counting the number of records in $Y$ that harbor $d_m$ exactly $c_m$ times and assigning these records to $r_m$ (steps 6-8). Then, in steps 11-14, GCCens determines the ICD code $d_o$ that requires the least amount of censoring and removes one instance of it from all records in $r_o$. To minimize the number of ICD codes that need to be modified, we remove one instance of $d_o$ per iteration. Subsequently, the cap $c_o$ for $d_o$ is decremented by 1 (step 15). Finally, GCCens releases a sample that satisfies $k$-map in step 17.

*Example* 1. Consider applying GCCens to the records in Figure 1a. We assume $k = 2$ and the cap set for $U = \{250, 272, 401, 724\}$ is $C = \{2, 2, 0, 1\}$. First, $Y$ is set equivalent to $S$ because no ICD code appears more than its cap. Next, GCCens finds that 2-map is not satisfied because $dis(D_3 = \{250, 250, 272\}) = 1$. So, GCCens censors one occurrence of 250 from $y_3$ (i.e., 250 is selected because its cap value is 2 and only one record in $Y$ has 2 instances of 250). After censoring, the cap value for 250 is decremented by 1, so $C = \{1, 2, 0, 1\}$. At this point, GCCens finds that 2-map is still not satisfied because $dis(D_2 = \{272, 272, 724\}) = 1$. Thus, one occurrence of 272 is censored from $y_2$. Finally, $Y$ satisfies 2-map and GCCens terminates. The resulting solution is depicted in Figure 1c.

### III.6 Data Utility Measure

When ICD code repeats are censored, there is a decrease in the utility of the data. To measure this utility loss, we introduce a measure called *Censoring Utility Loss* (CUL). This is defined as the number of censored ICD codes in a record $s$ divided by the total number of ICD codes in $s$. As an example, assume that we remove one instance of 272 from the second record in Figure 1a. In this case, CUL equals $1/3$ because there were three ICD codes in this record, one of which is censored. We note that the greedy heuristic in GCCens is designed to minimize the sum of CUL values in each iteration by choosing to censor the ICD code that incurs the minimum utility loss (see step 11 in Algorithm 1). The CUL values for the records in Figure 1c, computed with respect to the anonymization in Example 1, are depicted in Figure 1d.

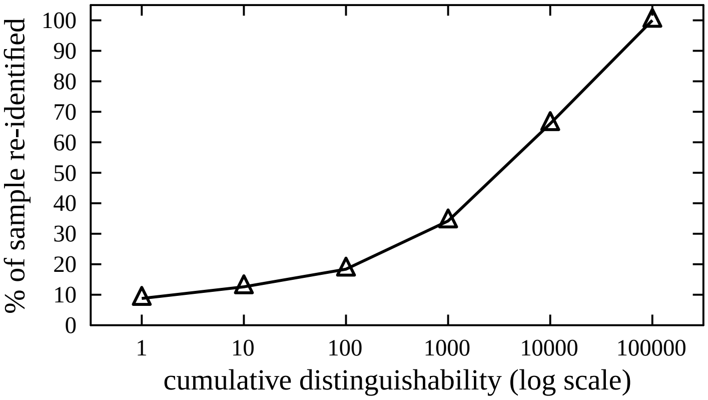


*Figure 2:* Distinguishability of the original patient records in the sample. A distinguishability of 1 means that a patient is uniquely identifiable.

## III.7 Experimental Evaluation

### III.7.1 Risk of Re-identification

Figure 2 summarizes the risk of associating a patient's record from the de-identified sample to their corresponding record in the population. This figure is a cumulative distribution and depicts the percent of patients in the sample (y-axis) that have a distinguishability score of a particular value or less (x-axis) with respect to the population from which they were derived. As can be seen, approximately 9% of the patients contained in the sample would be uniquely identifiable if the original data were disclosed. This confirms that a re-identification attack is feasible in practice and there is a need for developing a formal protection method.

### III.7.2 Effect of $k$ on Data Utility

Next, we recognize that not all data recipients will be comfortable working with a sample that is capped to varying degrees. Thus, we evaluated the effectiveness of GCCens in preserving utility when it is applied with all caps set to 3 (i.e., $C = \{3,3,3,3\}$) and various $k$ values between 5 and 25. Table 1 reports the mean, standard deviation, median, and skewness[1] of the distribution of the CUL values for all records in the sample. As expected, as we increase $k$ we find an increase in the mean of the CUL distribution. This is because GCCens needs to censor a larger number of ICD codes to meet a stricter privacy requirement. However, it is notable that GCCens retained 95.4% of the ICD codes on average when $k = 5$ as is often applied in practice (i.e., the mean of the CUL distribution was 0.046) [23]. We note that while 4.6% of the

[1] Skewness is a standard measure of the asymmetry of the distribution [53].

ICD codes in a record were censored on average, GCCens only modified 16% of the records in the sample. We also observed a positive skew in the CUL distribution for all tested values of $k$, which implies that the number of censored codes is closer to 0 for most patient records.

*Table 1:* Statistics on the distribution of CUL when GCCens was applied with all caps set to 3.

| $k$ | Mean | Std. Dev. | Median | Skewness |
|---|---|---|---|---|
| 5 | 0.046 | 0.123 | 0 | 3.016 |
| 10 | 0.046 | 0.123 | 0 | 3.016 |
| 25 | 0.091 | 0.156 | 0 | 1.501 |

### III.7.3 Effect of $C$ on Data Utility

Finally, we evaluated the impact of forcing all values in $C$ to be equivalent. For this set of experiments, we fixed $k$ to 5 and varied the cap between 3 and 10. The results are summarized in Table 2. Notice that GCCens performed a greater amount of censoring when larger cap values are supplied. This is expected because large cap values permit more information to be released, which makes it more difficult to generate a sufficient privacy solution [22, 28, 52]. However, GCCens managed to retain a reasonably large percentage of the ICD codes in all tested cases. In particular, 92% of the ICD codes were retained when the cap was set to 4 (i.e., the mean of the CUL distribution was 0.08). We believe this result is promising because the data derived by GCCens in this experiment was deemed to be useful for comorbidity analysis by a clinician. Moreover, when releasing at most 5 repeats of the ICD codes, GCCens retained, on average, 88.1% of the codes. We also observed a positive skew in the CUL distribution for all tested cap values, which implies that the number of censored codes is closer to 0 for most patient records.

## III.8 Summary

This chapter considered repeated data, a special case of more complex longitudinal data. Specifically, we first demonstrated the feasibility of a re-identification attack based on repeated diagnoses derived from real patient-specific clinical data. We then developed an algorithm to provide formal computational guarantees against such attacks. Our experiments verify that the proposed approach permits privacy-preserving patient record dissemination while retaining much of the information of the original records.

*Table 2:* Distribution statistics of CUL when GCCens was applied with $k = 5$ and all caps set to a particular value.

| Cap | Mean | Std. Dev. | Median | Skewness |
|---|---|---|---|---|
| 3 | 0.046 | 0.123 | 0 | 3.016 |
| 4 | 0.080 | 0.152 | 0 | 2.042 |
| 5 | 0.119 | 0.183 | 0 | 1.383 |
| 6 | 0.141 | 0.209 | 0 | 1.131 |
| 7 | 0.156 | 0.229 | 0 | 1.050 |
| 8 | 0.191 | 0.270 | 0 | 0.952 |
| 9 | 0.197 | 0.279 | 0 | 0.939 |
| 10 | 0.213 | 0.282 | 0 | 0.898 |

# CHAPTER IV

# ANONYMIZATION OF LONGITUDINAL DATA DERIVED FROM ELECTRONIC MEDICAL RECORDS

In Chapter III, we performed a pilot study on repeated data, which are a special case of more complex longitudinal data, and proposed a methodology to anonymize such data. In this chapter, we focus on the broader problem and present *Longitudinal Data Anonymizer* (LDA) which is the first approach to formally anonymize EMR-derived longitudinal data. Specifically, we first demonstrate the privacy problem, and formalize the notions of privacy and utility. We then present LDA and a baseline comparison algorithm. We conclude the chapter with an experimental evaluation of the proposed approach using several patient cohorts derived from the VUMC EMR system.

| $T$ | (ICD, Age) | DNA |
|---|---|---|
| *1* | (401.1, 33) (401.1, 34) (401.1, 35) | CT...A |
| *2* | (401.1, 38) (401.1, 40) | GC..A |
| *3* | (401.9, 38) (401.1, 40) | AC..A |
| *4* | (401.9, 33) (401.1, 33) (401.1, 34) (401.1, 35) | CC..A |
| *5* | (401.1, 39) (401.9, 40) | GC..C |
| *6* | (401.1, 40) (401.9, 40) | TG..A |

**(a)**

| ID | Name | YOB | Service Date | ICD |
|---|---|---|---|---|
| 1 | Tom | 1975 | 02/03/2008 | 401.1 |
| 1 | Tom | 1975 | 02/23/2009 | 401.1 |
| 1 | Tom | 1975 | 02/05/2010 | 401.1 |
| 2 | Jane | 1968 | 07/17/2006 | 401.1 |
| 2 | Jane | 1968 | 03/03/2008 | 401.1 |
| ... | ... | ... | ... | ... |
| 6 | Jim | 1966 | 07/02/2006 | 401.9 |

**(b)**

| $\check{T}$ | (ICD, Age) | DNA |
|---|---|---|
| *1* | (401.1, 33) (401.1, 34) (401.1, 35) | CT...A |
| *2* | (401, 38) (401.1, 40) | GC..A |
| *3* | (401, 38) (401.1, 40) | AC..A |
| *4* | (401.1, 33) (401.1, 34) (401.1, 35) | CC..A |
| *5* | (401.1, [39-40]) (401.9, 40) | GC..C |
| *6* | (401.1, [39-40]) (401.9, 40) | TG..A |

**(c)**

*Figure 3:* A depiction of the longitudinal data privacy problem. (a) and (b) depict longitudinal data and identified EMR, respectively. A 2-anonymization based on the proposed approach is depicted in (c).

## IV.1 Motivating Example

As an example of the problem studied in this chapter, consider the longitudinal data in Figure 3a. Each record corresponds to a fictional de-identified patient and is comprised of ICD codes, patient's age when a code was received, and a DNA sequence. For instance, the second record denotes that a patient was diagnosed with *benign essential hypertension* (code 401.1) at ages 38 and 40 and has the DNA sequence 'GC...A'. The clinical and genomic data are derived from an EMR system and a research project beyond primary care (i.e., they are not contained in the EMR system), respectively. Publishing the data of Figure

3a could allow a hospital employee with access to the EMR to associate *Jane* with her DNA sequence. This is because the identified record, shown in Figure 3b, can only be linked to the second record in Figure 3a based on the ICD code 401.1 and ages 38 and 40.

## IV.2 Background and Problem Formulation

This section begins with a high-level overview of the proposed approach. Next, we present the notation and the definitions for the privacy and adversarial models, the data transformation strategies, and the information loss metrics. We conclude the section with a formal problem description.

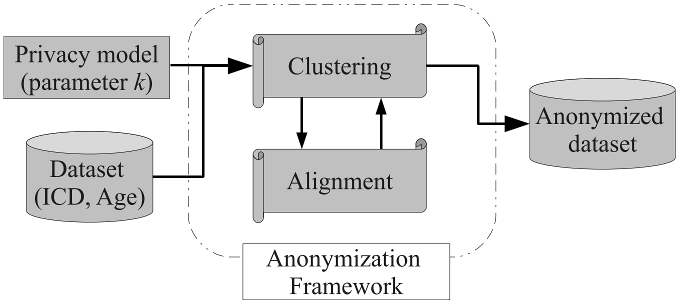


*Figure 4:* A general architecture of the longitudinal data anonymization process.

### IV.2.1 Architectural Overview

Figure 4 provides an overview of the data anonymization process. The process is initiated when the data owner supplies the following information: (1) a dataset of longitudinal patient records, each of which consists of (ICD, Age) pairs and a DNA sequence and (2) a parameter *k* that expresses the desired level of privacy. Given this information, the process invokes our anonymization framework. To satisfy the *k*-anonymity principle, our framework forms clusters of at least *k* records of the original dataset, which are modified using generalization and suppression.

### IV.2.2 Notation

A dataset $D$ consists of longitudinal records of the form $<T, DNA_T>$, where $T$ is a *trajectory*[1] and $DNA_T$ is a genomic sequence. Each trajectory corresponds to a distinct patient in $D$ and is a multiset[2] of pairs (i.e., $T = \{t_1, ..., t_m\}$) drawn from two attributes, namely ICD and Age (i.e., $t_i = (u \in \text{ICD}, v \in \text{Age})$), which contain the diagnosis codes assigned to a patient and their age, respectively. $|D|$ denotes the number of records in $D$ and $|T|$ the *length* of $T$, defined as the number of pairs in $T$. We use the '.' operator to refer to a specific attribute value in a pair (e.g., $t_i.icd$ or $t_i.age$). To study the data temporally, we order the pairs in $T$ with respect to Age, such that $t_{i-1}.age \leq t_i.age$.

### IV.2.3 Adversarial Model

We assume an adversary has access to the original dataset $D$, such as in Figure 3a. An adversary may perform a re-identification attack in several ways, such as:

- *Using identified EMR data:* The adversary links $D$ with the identified EMR data, such as those of Figure 3b, based on (ICD, Age) pairs. This scenario requires the adversary to have access to the identified EMR data, which is the case of an employee of the institution from which the longitudinal data were derived.

- *Using publicly available hospital discharge summaries and identified resources:* The adversary first links $D$ with hospital discharge summaries based on (ICD, Age) pairs to associate patients with certain demographics. In turn, these demographics are exploited in another linkage with public records, such as voter registration lists, which contain identity information [21, 26].

Note that in both cases, an adversary is able to link patients to their DNA sequences, which suggests a formal approach to longitudinal data anonymization is desirable.

### IV.2.4 Privacy Model

The formal definition of $k$-anonymity in the longitudinal data context is provided in Definition 1. Since each trajectory often contains multiple (ICD, Age) pairs, it is difficult to know which can be used by an adversary

[1] We use the term trajectory since the diagnosis codes at different ages can be seen as a route for the patient throughout his life.

[2] Contrary to a set, a multiset can contain an element more than once.

to perform re-identification attacks. Thus, we consider the worst-case scenario in which *any* combination of (ICD, Age) pairs can be exploited. Regardless, $k$-anonymity limits an adversary's ability to perform re-identification based on (ICD, Age) pairs, because each trajectory is associated with no less than $k$ patients.

*Definition* 1. ($k$-Anonymity) An anonymized dataset $\tilde{D}$, produced from $D$, is $k$-anonymous if each trajectory in $\tilde{D}$, projected over QI, appears at least $k$ times for any QI in $D$.

### IV.2.5 Data Transformation Strategies

Generalization and suppression are typically guided by a domain generalization hierarchy (Definition 2) [54].

*Definition* 2. (Domain Generalization Hierarchy) A domain generalization hierarchy (DGH) for attribute $\mathcal{A}$, referred to as $H_{\mathcal{A}}$, is a partially ordered tree structure which defines valid mappings between specific and generalized values of $\mathcal{A}$. The root of $H_{\mathcal{A}}$ is the most generalized value of $\mathcal{A}$, and is returned by a function $root$.

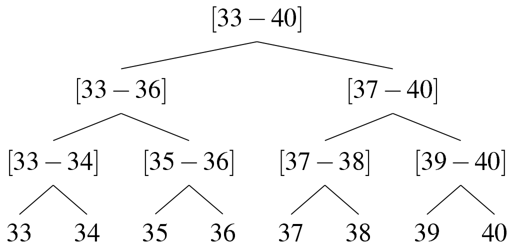


*Figure 5:* An example of the domain generalization hierarchy for Age.

*Example* 2. Consider $H_{Age}$ in Figure 5. The values in the domain of Age (i.e., 33, 34, ..., 40) form the leaves of $H_{Age}$. These values are then mapped to two, to four, and eventually to eight-year intervals. The root of $H_{Age}$ is returned by $root(H_{Age})$ as $[33-40]$.

Our approach does not impose any constraints on the structure of an attribute's DGH, such that the data owners have complete freedom in its design. For instance, for ICD codes, data owners can use the standard ICD-9-CM hierarchy.[3] For ages, data owners can use a pre-defined hierarchy (e.g., the age hierarchy in the HIPAA Safe Harbor Policy[4]) or design a DGH manually.[5]

[3] More information is available at `http://www.cdc.gov/nchs/icd.htm`

[4] The Safe Harbor standard of the HIPAA Privacy Rule states all ages under 89 can be retained intact, while 90 or greater must be grouped together. More information is available at http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/

[5] We further note that our approach can be extended to other categorical attributes, such as SNOMED-CT and Date, provided that a DGH can be specified for each of the attributes. Such extensions, however, are beyond the scope of this thesis.

According to Definition 3, each specific value of an attribute generalizes to its direct ancestor in a DGH. However, a specific value can be projected up multiple levels in a DGH via a sequence of generalizations. As a result, a generalized value $\mathcal{A}_i$ is interpreted as any one of the leaf nodes in the subtree rooted by $\mathcal{A}_i$ in $H_{\mathcal{A}}$.

*Definition* 3. (Generalization and Suppression) Given a node $\mathcal{A}_i \neq root(H_{\mathcal{A}})$ in $H_{\mathcal{A}}$, generalization is performed using a function $f{:}\mathcal{A}_i \rightarrow \mathcal{A}_j$ which replaces $\mathcal{A}_i$ with its direct ancestor $\mathcal{A}_j$. Suppression is a special case of generalization and is performed using a function $g{:}\mathcal{A}_i \rightarrow \mathcal{A}_r$ which replaces $\mathcal{A}_i$ with $root(H_{\mathcal{A}})$.

*Example* 3. Consider the last trajectory in Figure 3c. The first pair $(401.1, [39-40])$ is interpreted as either $(401.1, 39)$ or $(401.1, 40)$.

### IV.2.6 Information Loss

Generalization and suppression incur information loss because values are replaced by more general ones or eliminated. To capture the amount of information loss incurred by these operations, we quantify the normalized loss for each ICD code and Age value in a pair based on the Loss Metric (LM) (Definition 4) [43].

*Definition* 4. (Loss Metric) The information loss incurred by replacing a node $\mathcal{A}_i$ with its ancestor $\mathcal{A}_j$ in $H_{\mathcal{A}}$ is:

$$LM(\mathcal{A}_i, \mathcal{A}_j) = \frac{\mathcal{A}_j^{\triangle} - \mathcal{A}_i^{\triangle}}{|\mathcal{A}|}$$

where $\mathcal{A}_i^{\triangle}$ and $\mathcal{A}_j^{\triangle}$ denote the number of leaf nodes in the subtree rooted by $\mathcal{A}_i$ and $\mathcal{A}_j$ in $H_{\mathcal{A}}$, respectively, and $|\mathcal{A}|$ denotes the domain size of attribute $\mathcal{A}$.

*Example* 4. Consider $H_{Age}$ in Figure 5. The information loss incurred by generalizing $[33-34]$ to $[33-36]$ is $\frac{4-2}{8} = 0.25$ because the leaf-level descendants of $[33-34]$ are 33 and 34, those of $[33-36]$ are 33, 34, 35 and 36, and the domain of Age consists of the values 33 to 40.

To introduce the combined LM, which captures the total LM of replacing two nodes with their ancestor, provided in Definition 6, we use the notation of lowest common ancestor, provided in Definition 5.

*Definition* 5. (Lowest Common Ancestor) The lowest common ancestor (LCA) $\mathcal{A}_\ell$ of nodes $\mathcal{A}_i$ and $\mathcal{A}_j$ in $H_{\mathcal{A}}$ is the farthest node (in terms of height) from $root(H_{\mathcal{A}})$ such that (1) $\mathcal{A}_i = \mathcal{A}_\ell \texttt{ or } f^n(\mathcal{A}_i) = \mathcal{A}_\ell$ and (2) $\mathcal{A}_j = \mathcal{A}_\ell \texttt{ or } f^m(\mathcal{A}_j) = \mathcal{A}_\ell$, and is returned by a function $lca$.

*Definition* 6. (Combined Loss Metric) The combined LM of replacing nodes $\mathcal{A}_i$ and $\mathcal{A}_j$ with their LCA $\mathcal{A}_\ell$ is:

$$LM(\mathcal{A}_i + \mathcal{A}_j, \mathcal{A}_\ell) = LM(\mathcal{A}_i, \mathcal{A}_\ell) + LM(\mathcal{A}_j, \mathcal{A}_\ell)$$

Next, we define the LM for an anonymized trajectory (Definition 7) and dataset (Definition 8), which we keep separate for each attribute.

*Definition* 7. (Loss Metric for an Anonymized Trajectory) Given an anonymized trajectory $\tilde{T}$ and an attribute $\mathcal{A}$, the LM with respect to $\mathcal{A}$ is computed as:

$$LM(\tilde{T}, \mathcal{A}) = \sum_{i=1}^{|\tilde{T}|} LM(\tilde{t}_i.\mathcal{A}, \tilde{t}_i^*.\mathcal{A})$$

where $\tilde{t}_i^*.\mathcal{A}$ denotes the value $\tilde{t}_i.\mathcal{A}$ is replaced with.

*Definition* 8. (Loss Metric for an Anonymized Dataset) Given an anonymized dataset $\tilde{D}$ and an attribute $\mathcal{A}$, the LM with respect to attribute $\mathcal{A}$ is computed as:

$$LM(\tilde{D}, \mathcal{A}) = \frac{1}{|\tilde{D}|} \sum_{\tilde{T} \in \tilde{D}} \frac{LM(\tilde{T}, \mathcal{A})}{|\tilde{T}|}$$

For clarity, we refer to LM for the attributes ICD and Age using ILM and ALM, respectively (e.g., we use $ILM(\tilde{D})$ instead of $LM(\tilde{D}, \text{ICD})$).

### IV.2.7 Problem Statement

The longitudinal data anonymization problem is formally defined as follows.

*Problem:* Given a longitudinal dataset $D$, a privacy parameter $k$, and DGHs for attributes ICD and Age, construct an anonymized dataset $\tilde{D}$, such that (i) $\tilde{D}$ is $k$-anonymous, (ii) the order of the pairs in each trajectory of $D$ is preserved in $\tilde{D}$, and (iii) $ILM(\tilde{D}) + ALM(\tilde{D})$ is minimized.

## IV.3 Anonymization Framework

In this section, we present our framework for longitudinal data anonymization.

Many clustering algorithms can be applied to produce $k$-anonymous data [55, 56]. This involves organizing records into clusters of size at least $k$, which are anonymized together. In the context of longitudinal data, the challenge is to define a distance metric for trajectories such that a clustering algorithm groups *similar* trajectories. We define the distance between two trajectories as the cost (i.e., incurred information loss) of their anonymization as defined by the LM. The problem then reduces to finding an anonymized version $\tilde{T}$ of two given trajectories such that $ILM(\tilde{T}) + ALM(\tilde{T})$ is minimized.

Finding an anonymization of two trajectories can be achieved by finding a matching between the pairs of trajectories that minimizes their cost of anonymization. This problem, which is commonly referred to as sequence *alignment*, has been extensively studied in various domains, notably for the alignment of DNA sequences to identify regions of similarity in a way that the total pairwise edit distance between the sequences is minimized [57, 58].

To solve the longitudinal data anonymization problem, we propose *Longitudinal Data Anonymizer* (LDA), a framework that incorporates alignment and clustering as separate components, as shown in Figure 4. The objective of each component is summarized below:

1. *Alignment* attempts to find a minimal cost pair matching between two trajectories, and

2. *Clustering* interacts with the Alignment component to create clusters of at least $k$ records.

Next, we examine each component in detail and develop methodologies to achieve their objectives.

### IV.3.1 Alignment

There are no directly comparable approaches to the method we developed in this chapter. So, we introduce a simple heuristic, called *Baseline*, to establish a minimum performance benchmark for comparison purposes. Given trajectories $X = \{x_1,...,x_m\}$ and $Y = \{y_1,...,y_n\}$, $ILM(X)$ and $ALM(X)$, and DGHs $H_{ICD}$ and $H_{Age}$, Baseline aligns $X$ and $Y$ by matching their pairs on the same index.[6]

The pseudocode for Baseline is provided in Algorithm 2. This algorithm initializes an empty trajectory $\tilde{T}$ to hold the output of the alignment and then assigns $ILM(X)$ and $ALM(X)$ to variables $i$ and $a$, respectively (steps 1-2). Then, it determines the length of the shorter trajectory (step 3) and performs pair matching (steps 4-9). Specifically, for the pairs of the trajectories that have the same index, Baseline constructs a pair containing the LCAs of the ICD codes and Age values in these pairs (step 5), appends the constructed pair to $\tilde{T}$ (step 6), and updates $i$ and $a$ with the information loss incurred by the generalizations (steps $7-8$). Next, Baseline updates $i$ and $a$ with the amount of information loss incurred by suppressing the ICD codes

[6] $ILM(X)$ and $ALM(X)$ are provided as input because $X$ may already be an anonymized version of two other trajectories.

**Algorithm 2** Baseline($X$, $Y$)

```
Require: Trajectories $X = \{x_1, ..., x_m\}$ and $Y = \{y_1, ..., y_n\}$, $ILM(X)$ and $ALM(X)$, DGHs $H_{ICD}$ and $H_{Age}$
Return: Anonymized trajectory $\tilde{T}$, $ILM(\tilde{T})$ and $ALM(\tilde{T})$
 1: $\tilde{T} \leftarrow \emptyset$
 2: $i \leftarrow ILM(X)$, $a \leftarrow ALM(X)$
 3: $s \leftarrow$ the length of the shorter of $X$ and $Y$
 4: for all $j \in [1 - s]$ do
        ▷Construct a pair containing the LCAs of $x_j$ and $y_j$
 5:     $p \leftarrow (lca(x_j.icd, y_j.icd, H_{ICD}), lca(x_j.age, y_j.age, H_{Age}))$
        ▷Append the constructed pair to $\tilde{T}$
 6:     $\tilde{T} \leftarrow \tilde{T} \cup p$
        ▷Information loss incurred by generalizing $x_j$ with $y_j$
 7:     $i \leftarrow i + ILM(x_j + y_j, p.icd)$
 8:     $a \leftarrow a + ALM(x_j + y_j, p.age)$
 9: end for
10: $Z \leftarrow$ the longer of $X$ and $Y$
11: for all $j \in [(s+1) - |Z|]$ do
        ▷Information loss incurred by suppressing $z_j$
12:     $i \leftarrow i + ILM(z_j, root(H_{ICD}))$
13:     $a \leftarrow a + ALM(z_j, root(H_{Age}))$
14: end for
15: return $\{\tilde{T}, i, a\}$
```

and Age values from the unmatched pairs in the longer trajectory (steps 10-14). Last, this algorithm returns $\tilde{T}$ along with $i$ and $a$, which correspond to $ILM(\tilde{T})$ and $ALM(\tilde{T})$, respectively (step 15).

To help preserve data utility, we provide *Alignment using Generalization and Suppression* (A-GS), an algorithm that uses dynamic programming to construct an anonymized trajectory that incurs minimal cost.

Before discussing A-GS, we briefly discuss the application of dynamic programming. The latter technique can be used to solve problems based on combining the solutions to subproblems which are not independent and share subsubproblems [59]. A dynamic programming algorithm stores the solution of a subsubproblem in a table to which it refers every time the subsubproblem is encountered. To give an example, for trajectories $X = \{x_1, ..., x_m\}$ and $Y = \{y_1, ..., y_n\}$, a subproblem may be to find a minimal cost pair matching between the first to the $j$-th pairs. A solution to this subproblem can be determined using solutions for the following subsubproblems and applying the respective operations:

- Align $X = \{x_1, ..., x_{j-1}\}$ and $Y = \{y_1, ..., y_{j-1}\}$, and generalize $x_j$ with $y_j$
- Align $X = \{x_1, ..., x_{j-1}\}$ and $Y = \{y_1, ..., y_j\}$, and suppress $x_j$
- Align $X = \{x_1, ..., x_j\}$ and $Y = \{y_1, ..., y_{j-1}\}$, and suppress $y_j$

Each case is associated with a cost. Our objective is to find an anonymized trajectory $\tilde{T}$, such that $ILM(\tilde{T}) + ALM(\tilde{T})$ is minimized, so we examine each possible solution and select the one with minimum information loss.

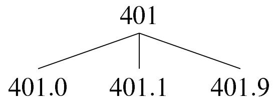


*Figure 6:* An example of the hypertension subtree in the ICD domain generalization hierarchy.

For a more specific example, consider the alignment of $X = \{(401.1,34)\ (401.1,35)\ (401.1,37)\}$ and $Y = \{(401.1,34)\ (401.1,36)\}$ using the DGHs shown in Figures 5 and 6. A solution for this problem can be determined using solutions for the following subproblems and applying the respective operations:

- Align $X = \{(401.1,34)\ (401.1,35)\}$ and $Y = \{(401.1,34)\}$, and generalize $(401.1,37)$ with $(401.1,36)$
- Align $X = \{(401.1,34)\ (401.1,35)\}$ and $Y = \{(401.1,34)\ (401.1,36)\}$, and suppress $(401.1,37)$
- Align $X = \{(401.1,34)\ (401.1,35)\ (401.1,37)\}$ and $Y = \{(401.1,34)\}$, and suppress $(401.1,36)$

The solution for the first subproblem is $\tilde{T} = \{(401.1,34)\}$. The second pair of $X$ (i.e., $(401.1,35)$) is suppressed, thus, the ILM and ALM associated with this solution are both 1. Furthermore, the first solution specifies that the last pair of $X$ and $Y$ (i.e., $(401.1,37)$ and $(401.1,36)$) are generalized together. The ILM associated with this operation is 0 as the pairs contain the same ICD code. According to the DGH for Age in Figure 5, the LCA of 36 and 37 is $[33-40]$, thus, the ALM associated with this operation is $1 * 2 = 2$. Therefore, the total LM for the first solution is 4. The solution for the second subproblem is $\tilde{T} = \{(401.1,34)\ (401.1,[35-36])\}$[7] and this solution is associated with an ILM of 0 and ALM of $\frac{2}{8} * 2 = 0.5$. Furthermore, the second solution specifies that the last pair of $X$ (i.e., $(401.1,37)$) is suppressed. The ILM and ALM associated with this operation are both 1. Therefore, the total LM for the second solution is 2.5. Similarly, the total LM for the third solution is 6. The solution with minimum information loss is the second one, thus, the alignment of $X$ and $Y$ is determined as $\tilde{T} = \{(401.1,34)\ (401.1,[35-36])\}$.

A-GS uses a similar approach to align trajectories. The algorithm accepts the same inputs, as well as weights $w_{ICD}$ and $w_{Age}$. The weights allow A-GS to control the information loss incurred by anonymizing

[7] This is because when this subproblem is divided into its subsubproblems, this is the solution with minimum information loss.

the values of each attribute. The data owners specify the attribute weights such that $w_{ICD} \geq 0$, $w_{Age} \geq 0$ and $w_{ICD} + w_{Age} = 1$. The pseudocode for A-GS is provided in Algorithm 3.

In step 1, A-GS initializes three matrices; $i$, $a$ and $r$. The first row (index 0) of each of these matrices corresponds to a *null* value, and starting from index 1, each row corresponds to a value in $X$. Similarly, the first column (indexed 0) of each of these matrices corresponds to a *null* value, and starting from index 1, each column corresponds to a value in $Y$. Specifically, for indices $h$ and $j$, $r_{h,j}$ records which of the following operations incurs minimum information loss: (i) generalizing $x_h$ and $y_j$ (denoted with $<\nwarrow>$), (ii) suppressing $x_h$ (denoted with $<\uparrow>$), and (iii) suppressing $y_j$ (denoted with $<\leftarrow>$). The entries in $i_{h,j}$ and $a_{h,j}$ keep the total ILM and ALM for aligning the subtrajectories $X_{sub} = \{x_1, ..., x_h\}$ and $Y_{sub} = \{y_1, ..., y_j\}$, respectively.

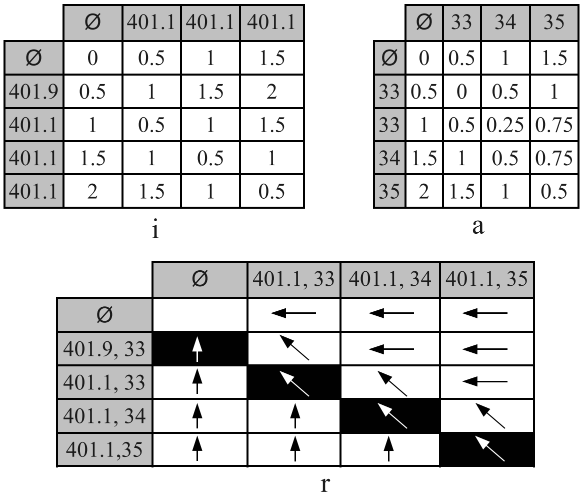
In step 2, A-GS assigns $ILM(X)$ and $ALM(X)$ to $i_{0,0}$ and $a_{0,0}$, respectively. We include *null* values in the rows and columns of $i$, $a$ and $r$ because at some point during alignment A-GS may need to suppress some portion of the trajectories. Therefore, in steps $3-7$ and $8-12$, A-GS initializes $i$, $a$ and $r$ for the values in $X$ and $Y$, respectively. Specifically, for indices $h$ and $j$, $i_{h,0}$ and $i_{0,j}$ keep the ILM for suppressing every pair in the subtrajectories $X_{sub} = \{x_1, ..., x_h\}$ and $Y_{sub} = \{y_1, ..., y_j\}$, respectively. Similar reasoning applies to matrix $a$. The first row and column of $r$ holds $<\uparrow>$ and $<\leftarrow>$ for suppressing values from $X$ and $Y$, respectively.

In steps $13-25$, A-GS performs dynamic programming. Specifically, for indices $h$ and $j$, A-GS determines a minimal cost pair matching of the subtrajectories $X_{sub} = \{x_1, ..., x_h\}$ and $Y_{sub} = \{y_1, ..., y_j\}$ based on the three cases listed above. Specifically, in steps $15-21$, A-GS constructs two temporary arrays, $c$ and $g$, to store the ILM and ALM for each possible solution, respectively. Next, in steps $22-23$, A-GS determines the solution with the minimum information loss and assigns the ILM, ALM and operation associated with the solution to $i_{h,j}$, $a_{h,j}$ and $r_{h,j}$, respectively. If there is a tie between the solutions, A-GS selects generalization as the operation for the sake of retaining more information.

In steps $26-36$, A-GS constructs the anonymized trajectory $\tilde{T}$ by traversing the matrix $r$. Specifically, for two pairs in the trajectories, if generalization incurs minimum information loss, A-GS appends to $\tilde{T}$ a pair containing the LCAs of the ICD codes and Age values in these pairs. The unmatched pairs in the trajectories are ignored during this process because A-GS suppresses these pairs. Finally, in step 37, Baseline returns $\tilde{T}$ along with $i_{m,n}$ and $a_{m,n}$, which correspond to $ILM(\tilde{T})$ and $ALM(\tilde{T})$, respectively.

**Algorithm 3** A-GS($X$, $Y$)

```
Require: Trajectories $X = \{x_1, ..., x_m\}$ and $Y = \{y_1, ..., y_n\}$, $ILM(X)$ and $ALM(X)$, DGHs $H_{ICD}$ and $H_{Age}$,
    weights $w_{ICD}$ and $w_{Age}$
Return: Anonymized trajectory $\tilde{T}$, $ILM(\tilde{T})$ and $ALM(\tilde{T})$
$\{i, a, r\} \leftarrow$ generate $(m+1) \times (n+1)$ matrices
$i_{0,0} \leftarrow ILM(X), a_{0,0} \leftarrow ALM(X)$
▷Initialize $i$, $a$ and $r$ with respect to $X$
for all $h \in [1-m]$ do
    $i_{h,0} \leftarrow i_{h-1,0} + ILM(x_h, root(H_{ICD})) \times w_{ICD}$
    $a_{h,0} \leftarrow a_{h-1,0} + ALM(x_h, root(H_{Age})) \times w_{Age}$
    $r_{h,0} \leftarrow <\uparrow>$
end for
▷Initialize $i$, $a$ and $r$ with respect to $Y$
for all $j \in [1-n]$ do
    $i_{0,j} \leftarrow i_{0,j-1} + ILM(y_j, root(H_{ICD})) \times w_{ICD}$
    $a_{0,j} \leftarrow a_{0,j-1} + ALM(y_j, root(H_{Age})) \times w_{Age}$
    $r_{0,j} \leftarrow <\leftarrow>$
end for
for all $h \in [1-m]$ do
    for all $j \in [1-n]$ do
        $\{c, g\} \leftarrow$ generate arrays with indices $<\nwarrow>, <\leftarrow>, <\uparrow>$
        ▷Compute the ILM for the possible solutions
        $c_{<\nwarrow>} \leftarrow i_{h-1,j-1} + ILM(x_h + y_j, lca(x_h.icd, y_j.icd, H_{ICD})) \times w_{ICD}$
        $c_{<\leftarrow>} \leftarrow i_{h,j-1} + ILM(y_j, root(H_{ICD})) \times w_{ICD}$
        $c_{<\uparrow>} \leftarrow i_{h-1,j} + ILM(x_h, root(H_{ICD})) \times w_{ICD}$
        ▷Compute the ALM for the possible solutions
        $g_{<\nwarrow>} \leftarrow a_{h-1,j-1} + ALM(x_h + y_j, lca(x_h.age, y_j.age, H_{Age})) \times w_{Age}$
        $g_{<\leftarrow>} \leftarrow a_{h,j-1} + ALM(y_j, root(H_{Age})) \times w_{Age}$
        $g_{<\uparrow>} \leftarrow a_{h-1,j} + ALM(x_h, root(H_{Age})) \times w_{Age}$
        ▷Solution with the minimum overall LM
        $w \leftarrow \text{argmin}_{u \in \{<\nwarrow>, <\leftarrow>, <\uparrow>\}} \{c_u + g_u\}$
        $i_{h,j} \leftarrow c_w$, $a_{h,j} \leftarrow g_w$, $r_{h,j} \leftarrow w$
    end for
end for
$\tilde{T} \leftarrow \emptyset$
$h \leftarrow m, j \leftarrow n$
▷Construct the anonymized trajectory $\tilde{T}$
while $h \geq 1$ or $j \geq 1$ do
    if $r_{h,j} = <\nwarrow>$ then
        $p \leftarrow (lca(x_h.icd, y_j.icd, H_{ICD}), lca(x_h.age, y_j.age, H_{Age}))$
        $\tilde{T} \leftarrow \tilde{T} \cup p$
        $h \leftarrow h-1$, $j \leftarrow j-1$
    end if
    if $r_{h,j} = <\leftarrow>$ then $j \leftarrow j-1$ end if
    if $r_{h,j} = <\uparrow>$ then $h \leftarrow h-1$ end if
end while
return $\{\tilde{T}, i_{m,n}, a_{m,n}\}$
```



| | Ø | 401.1 | 401.1 | 401.1 |
|---|---|---|---|---|
| Ø | 0 | 0.5 | 1 | 1.5 |
| 401.9 | 0.5 | 1 | 1.5 | 2 |
| 401.1 | 1 | 0.5 | 1 | 1.5 |
| 401.1 | 1.5 | 1 | 0.5 | 1 |
| 401.1 | 2 | 1.5 | 1 | 0.5 |

i

| | Ø | 33 | 34 | 35 |
|---|---|---|---|---|
| Ø | 0 | 0.5 | 1 | 1.5 |
| 33 | 0.5 | 0 | 0.5 | 1 |
| 33 | 1 | 0.5 | 0.25 | 0.75 |
| 34 | 1.5 | 1 | 0.5 | 0.75 |
| 35 | 2 | 1.5 | 1 | 0.5 |

a

| | Ø | 401.1, 33 | 401.1, 34 | 401.1, 35 |
|---|---|---|---|---|
| Ø | | ← | ← | ← |
| 401.9, 33 | ↑ | ↖ | ← | ← |
| 401.1, 33 | ↑ | ↖ | ↖ | ← |
| 401.1, 34 | ↑ | ↑ | ↖ | ↖ |
| 401.1,35 | ↑ | ↑ | ↑ | ↖ |

r

*Figure 7:* Matrices *i*, *a* and *r* for $T_1$ and $T_4$ in Figure 3a. The columns and rows of these matrices correspond to the values in $T_1$ and $T_4$, respectively. This alignment uses the domain generalization hierarchies in Figures 5 and 6, and assumes that $w_{ICD} = w_{Age} = 0.5$.

*Example* 5. Consider applying A-GS to $T_1$ and $T_4$ in Figure 3a using the DGHs shown in Figures 5 and 6 and assuming that $w_{ICD} = w_{Age} = 0.5$. The matrices *i*, *a* and *r* are illustrated in Figure 7. As $T_1$ and $T_4$ are not anonymized, we initialize $i_{0,0} = a_{0,0} = 0$. Subsequently, A-GS computes the values for the entries in the first row and column of the matrices. For instance, $i_{0,3}$ keeps the ILM for suppressing all ICD codes from $T_1$ and has a value of $1 + (1 * 0.5) = 1.5$. This is computed by summing the ILM for suppressing the first two ICD codes (i.e., the value stored in $i_{0,2}$) with the weight-adjusted ILM for suppressing the third ICD code. Then, A-GS performs dynamic programming. The process starts with aligning $T_{1,sub} = \{(401.1, 33)\}$ and $T_{4,sub} = \{(401.9, 33)\}$. The possible solutions for this subproblem are:

- Align $T_{1,sub} = \{\emptyset\}$ and $T_{4,sub} = \{\emptyset\}$, and generalize 401.1 with 401.9 and 33 with 33
- Align $T_{1,sub} = \{(401.1, 33)\}$ and $T_{4,sub} = \{\emptyset\}$, and suppress 401.9 and 33
- Align $T_{1,sub} = \{\emptyset\}$ and $T_{4,sub} = \{(401.9, 33)\}$, and suppress 401.1 and 33

The ILM and ALM for the subsubproblem in the first solution are stored in $i_{0,0}$ and $a_{0,0}$, respectively. Generalizing the ICD code 401.1 with the ICD code 401.9 has an ILM of $(1+1) * 0.5 = 1$, and generalizing the age 33 with the age 33 has an ALM of 0. Therefore, the first solution has a total LM of 1. The ILM and ALM for the subsubproblem in the second solution are stored in $i_{0,1}$ and $a_{0,1}$, respectively. The suppression of the ICD code 401.9 and the age 33 has an ILM and ALM of $1 * 0.5 = 0.5$. Therefore, the second solution has a total LM of 2. Similarly, the third solution has a total LM of 2. The solution with the minimum information loss is the first one, hence, A-GS stores 1, 0 and $<\nwarrow>$ in $i_{1,1}$, $a_{1,1}$ and $r_{1,1}$, respectively. After the values for the remaining entries are computed, A-GS uses the matrix $r$ to construct the anonymized trajectory $\tilde{T}$. The process starts with examining the bottom-right entry, which denotes a generalization. As a result, A-GS appends $(401.1, 35)$ to $\tilde{T}$. The process continues by following the symbols and A-GS returns $\tilde{T} = \{(401.1, 33), (401.1, 34), (401.1, 35)\}$ along with $i_{4,3}$ and $a_{4,3}$, which correspond to $ILM(\tilde{T})$ and $ALM(\tilde{T})$, respectively.

### IV.3.2 Clustering

We base our methodology for the clustering component on the Maximum Distance to Average Vector (MDAV) algorithm [60], an efficient heuristic for $k$-anonymity. The algorithm iteratively selects the most frequent trajectory in a longitudinal dataset, finds its most distant trajectory, and forms a cluster of at least $k$ records around the latter. We define the distance between two trajectories as the cost of their anonymization. As such, the most distant trajectory to a given trajectory is the one which maximizes the sum of ILM and ALM returned from A-GS.

A similar reasoning applies while we form a cluster, i.e., we choose to add the trajectory which minimizes the sum of ILM and ALM returned from A-GS. The clustering component returns $\tilde{D}$, a $k$-anonymized version of the longitudinal dataset, along with $ILM(\tilde{D})$ and $ALM(\tilde{D})$.

## IV.4 Experimental Evaluation

This section presents an experimental evaluation of the anonymization framework. We compare the anonymization methods on data utility, as indicated by the LM measure and aggregate query answering accuracy. Furthermore, we verify the ability of our approach to preserve the utility of certain values that are important for known biomedical analysis.

### IV.4.1 Experimental Setup and Metrics

We conducted experiments with three datasets derived from the Synthetic Derivative (SD), a collection of de-identified information extracted from the EMR system of the VUMC [50]. Figure 8 summarizes our

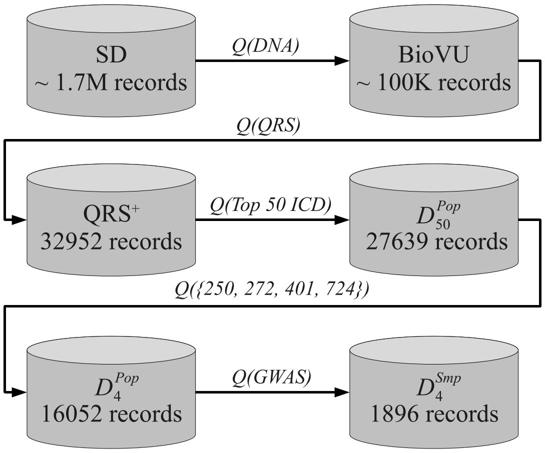


*Figure 8:* Dataset creation process. The function $Q$ is used to issue a query.

dataset creation process. We first issued a query to retrieve the records of patients whose DNA samples were genotyped and stored in BioVU, VUMC's DNA repository linked to the SD. Then, using the QRS phenotype specification in [61], we identified the patients eligible to participate in a GWAS on native electrical conduction within the ventricles of the heart. Subsequently, we created a dataset called $D_{50}^{Pop}$ by restricting our query to the 50 most frequent ICD codes that occur in at least 5% of the records in BioVU. Next, we created a dataset called $D_{4}^{Pop}$, which is a subset of $D_{50}^{Pop}$, containing the following comorbid ICD codes selected for Chapter III and [25]: *250* (diabetes mellitus), *272* (disorders of lipoid metabolism), *401* (essential hypertension), and *724* (other and unspecified disorders of the back). Finally, we created a dataset called $D_{4}^{Smp}$, which is a subset of $D_{4}^{Pop}$, containing the records of patients who actually participated in the aforementioned GWAS [62]. A variant of $D_{4}^{Smp}$ has been used in Chapter III and [25] with no temporal information and more records because some patients had invalid year of birth values in the SD, and thus, could not be included to $D_{4}^{Smp}$. We further note that $D_{4}^{Smp}$ is expected to be deposited into the dbGaP repository. The characteristics of our datasets are summarized in Table 3. Note that $D_{4}^{Pop}$ is a reduced version of $D_{50}^{Pop}$ in terms of domain size, and similarly, $D_{4}^{Smp}$ is a reduced version of $D_{4}^{Pop}$ in terms of dataset size. Hence, our datasets will allow us to capture the effects of domain and dataset size on anonymization.

*Table 3:* Descriptive summary statistics of the datasets.

| $D$ | $\lvert D\rvert$ | $\lvert$(ICD, Age)$\rvert$ | $\lvert$ICD$\rvert$ | $\lvert$Age$\rvert$ | Avg. (ICD, Age) per $T$ | Avg. ICD per $T$ | Avg. Age per $T$ |
|---|---|---|---|---|---|---|---|
| $D_{50}^{Pop}$ | 27639 | 4246 | 50 | 102 | 9.32 | 5.88 | 3.10 |
| $D_{4}^{Pop}$ | 16052 | 354 | 4 | 97 | 4.05 | 1.65 | 2.78 |
| $D_{4}^{Smp}$ | 1896 | 322 | 4 | 90 | 6.39 | 1.96 | 4.04 |

Throughout our experiments, we varied $k$ between 2 and 15, noting that $k = 5$ tends to be applied in practice [44]. Initially, we set $w_{ICD} = w_{Age} = 0.5$, and we measured the impact of varying these parameters in a later subsection. We implemented all algorithms in Java and conducted our experiments on an Intel 2.8GHz powered system with 4GB RAM.

To quantify information loss, we assumed a scenario in which a scientist issues queries on anonymized data to retrieve the number of trajectories that harbor a combination of (ICD, Age) pairs that appear in the original trajectories. Such queries are typical in many biomedical data mining applications [52]. To quantify the accuracy of answering such a workload of queries, we used the *Average Relative Error* (AvgRE) measure [39].

Given a workload of queries, the AvgRE captures the accuracy of answering these queries on an anonymized dataset. The queries we consider can be modeled as follows:[8]

```
Q: SELECT COUNT(*)
   FROM dataset
   WHERE (u ∈ ICD, v ∈ Age) ∈ dataset, ...
```

Let $a$(Q) be the answer of a COUNT() query Q when it is issued on the original dataset. The value of $a$(Q) can be easily obtained by counting the number of trajectories in the original dataset that contain the (ICD, Age) pairs in Q.

Let $e$(Q) be the answer of Q when it is issued on the anonymized dataset. This is an estimate because a generalized value is interpreted as any leaf node in the subtree rooted by that value in the DGH. Therefore, an anonymized pair may correspond to any pair of possible ICD codes and Age values, assuming each pair is equally likely. The value of $e$(Q) can be obtained by computing the probability that a trajectory in the anonymized dataset satisfies Q, and then summing these probabilities across all trajectories.

To illustrate how an estimate can be computed, assume that a data recipient issues a query for the number of patients diagnosed with ICD code 401.1 at age 39 using the anonymized dataset in Figure 3c. Referring to the DGHs in Figures 5 and 6, it can be seen that the only trajectories that may contain $(401.1, 39)$ are the last two since they contain the generalized pair $(401.1, [39-40])$. Furthermore, observe that 401.1 is a leaf node in Figure 6, hence the set of possible ICD codes is $\{401.1\}$. Similarly, the subtree rooted by $[39-40]$ in Figure 5 consists of two leaf nodes, hence the set of possible Age values is $\{39, 40\}$. Therefore, there are

[8] The queries we consider form the basis for other query types such as range queries which return the number trajectories that harbor a combination of ICD codes in a given range of Age values.

two possible pairs: $\{(401.1,39),(401.1,40)\}$, and the probability that one of the trajectories was originally harboring $(401.1,39)$ is $\frac{1}{2}$. Then, an approximate answer for the query is computed as $\frac{1}{2} \times 2 = 1$.

The *Relative Error* (RE) for an arbitrary query Q is computed as RE(Q) = $|a(\text{Q}) - e(\text{Q})|/a(\text{Q})$. For instance, the RE for the above example query is $|1-1|/1 = 0$ since the original dataset in Figure 3a contains one trajectory with $(401.1,39)$.

The AvgRE for a workload of queries is the mean RE of all issued queries. It reflects the mean error in answering the query workload.

### IV.4.2 Capturing Data Utility Using LM

We first compared the algorithms with respect to the LM.

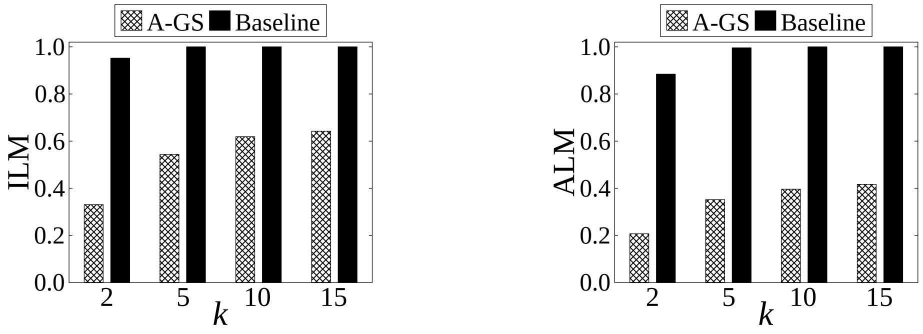


*Figure 9:* A comparison of information loss for $D_{50}^{Pop}$ using various $k$ values.

Figure 9 depicts the results with $D_{50}^{Pop}$. The ILM and ALM increase with $k$ for both algorithms, which is expected because as $k$ increases, a larger amount of distortion is needed to satisfy a stricter privacy requirement. Note that Baseline incurred substantially more information loss than A-GS for all $k$. In fact, Baseline failed to construct a practically useful result when $k > 2$, as it suppressed all values from the dataset.

Figure 10 shows the results of the same experiment performed on $D_4^{Pop}$. Observe that A-GS achieved a much better result than Baseline for all tested $k$ values. Interestingly, A-GS incurred less information loss on $D_4^{Pop}$ than $D_{50}^{Pop}$, which is important because a relatively small number of ICD codes may suffice to study a range of different diseases [11, 25].

The results shown in Figure 11 for $D_4^{Smp}$ are qualitatively similar to that of Figure 10. Again, A-GS significantly outperformed Baseline. It is also worthwhile to note that the information loss incurred by our approach remains relatively low (i.e., below 0.5), even though the ILM and ALM values are slightly larger

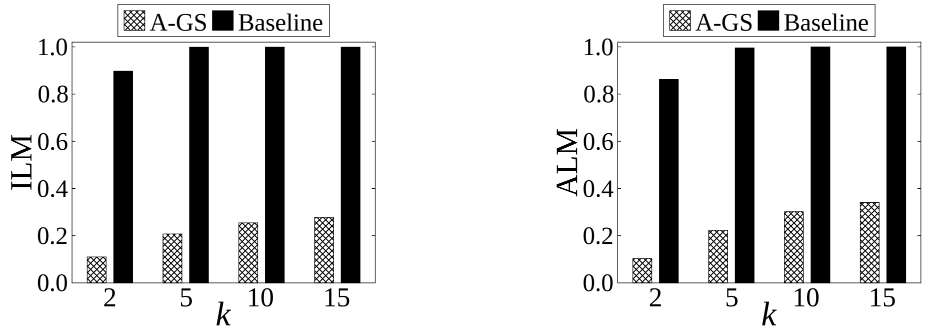


*Figure 10:* A comparison of information loss for $D_4^{Pop}$ using various $k$ values.

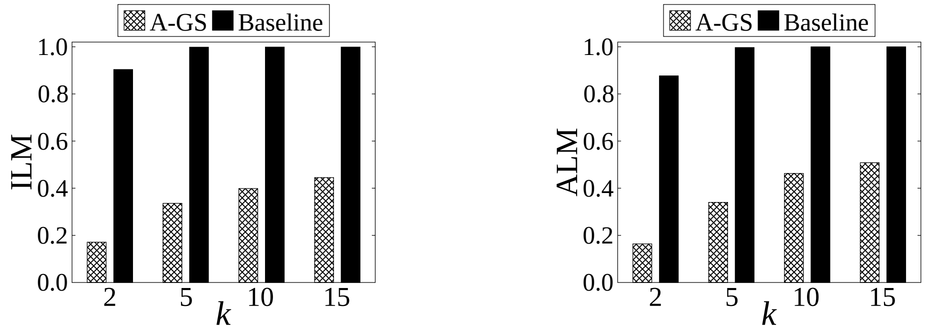


*Figure 11:* A comparison of information loss for $D_4^{Smp}$ using various $k$ values.

than those of Figure 10. This is attributed to the fact that $D_4^{Smp}$ is more sparse than $D_4^{Pop}$, which implies that it is more difficult to anonymize [63].

### IV.4.3 Capturing Data Utility Using AvgRE

We next analyzed the effectiveness of our approach for supporting general biomedical analysis using the AvgRE measure with a workload of queries that are comprised of the combination of (ICD, Age) pairs that appear in at least 1% of the original trajectories of a dataset.

Figure 12 shows the AvgRE scores of running A-GS on our datasets. The results for Baseline are not reported because they were more than 6 times worse than our approach for $k = 2$, and the worst possible for $k > 2$. This is because Baseline suppressed all values. As expected, we find an increase in AvgRE scores as $k$ increases, which is due to the privacy/utility tradeoff. Nonetheless, A-GS allows fairly accurate query answering on each dataset by having an AvgRE score of less than 1 for all $k$. The results suggest our approach can be effective, even when a high level of privacy is required. Furthermore, we observe that

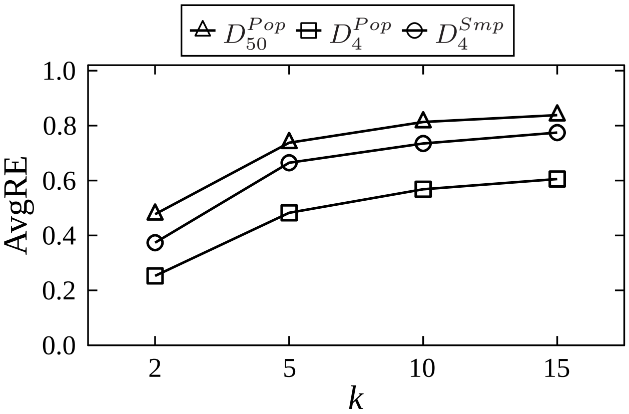


*Figure 12:* A comparison of query answering accuracy for $D_{50}^{Pop}$, $D_{4}^{Pop}$ and $D_{4}^{Smp}$ using various $k$ values.

the AvgRE scores for $D_{4}^{Pop}$ are lower than those of $D_{4}^{Smp}$, which are in turn lower than those of $D_{50}^{Pop}$. This implies that query answers are more accurate for small domain sizes and large datasets.

### IV.4.4 Capturing Data Utility Using Histograms

We also investigated how the distributions of ICD and Age are affected by anonymization. Figures 13 and 14 report the percent of ICD codes and Age values at each level of the DGHs after anonymizing $D_{50}^{Pop}$ and $D_{4}^{Pop}$ using $k = 5$, respectively. Note that the standard ICD-9-CM hierarchy consists of 5 levels (i.e., 5-digit ICD codes, 3-digit ICD codes, Sections, Chapters, and Any which corresponds to the least specific ICD code of $001 - 999$). The domain of Age consists of the values 1 to 128, and the DGH for Age is a 8-level full binary tree[9]. The results for Baseline are not reported because it suppressed all values in the datasets (i.e., all ICD codes and Age values appear in the Any and 128-year levels, respectively). For $D_{50}^{Pop}$, observe that A-GS retained more than 30% of the ICD codes at the 3-digit level or below. Similarly, more than 60% of the Age values are retained at the 16-year level or below. For $D_{4}^{Pop}$, we did not observe any ICD code at the 5-digit and Section levels because $D_{4}^{Pop}$ consists of 3-digit ICD codes and the LCA of any two ICD codes in this dataset is located at the Chapter and Any levels. As can be seen, A-GS retained more than 55% of the ICD codes at their original, 3-digit, level. Notably, more than 70% of the Age values are retained at the 16-year level or below. These results are promising as they suggest that the anonymized data derived by our approach may be utilized in longitudinal clinical investigations. We, however, note that further research is necessary to determine how the resulting anonymizations impact specific clinical phenotype definitions

[9] A full binary tree is a tree in which every node other than the leaves has two children [59].

of interest.

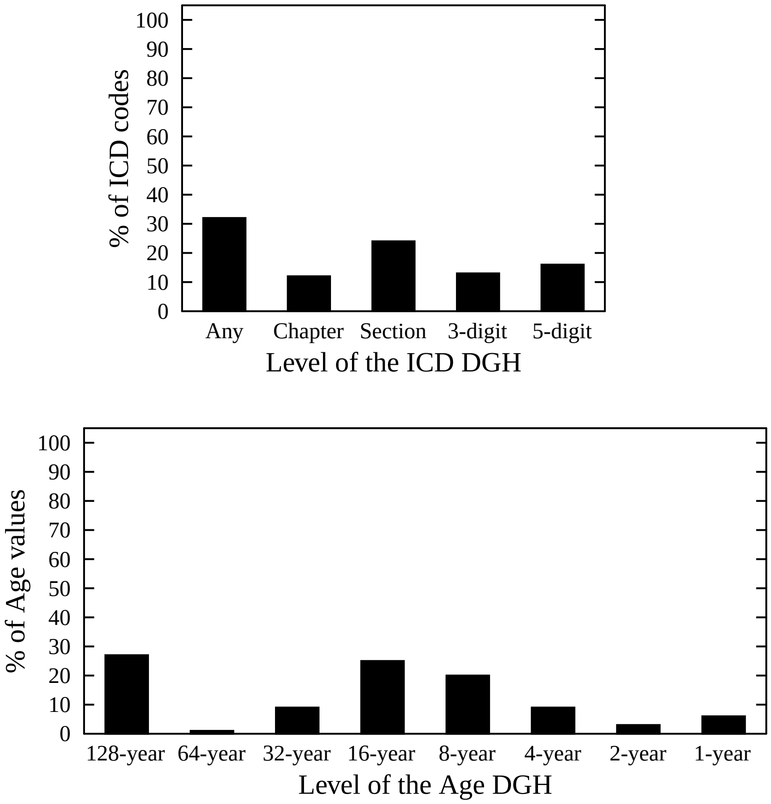


*Figure 13:* The histogram of ICD and Age after anonymizing $D_{50}^{Pop}$ using A-GS with $k = 5$.

### IV.4.5 Prioritizing Attributes

Next, we investigated how configurations of attribute weighting affect information loss. Figure 15 reports the results for $D_{50}^{Pop}$ and $k = 2$ when our algorithm is configured with weights ranging from 0.1 to 0.9. Observe that when $w_{ICD} = 0.1$ and $w_{Age} = 0.9$, A-GS distorted Age values much less than ICD values. Similarly, A-GS incurred less information loss for ICD than Age when we specified $w_{ICD} = 0.9$ and $w_{Age} = 0.1$. This result implies that data managers can use weights to bias utility towards either attribute.

### IV.4.6 Real World Anonymized Data

Finally, we analyzed some of the anonymized trajectories derived by our approach. Table 4 presents two clusters generated for $D_4^{Smp}$ using $k = 5$. These clusters are selected because they contain generalized and suppressed values (i.e., there are other clusters with less information loss). Note that the trajectories in the

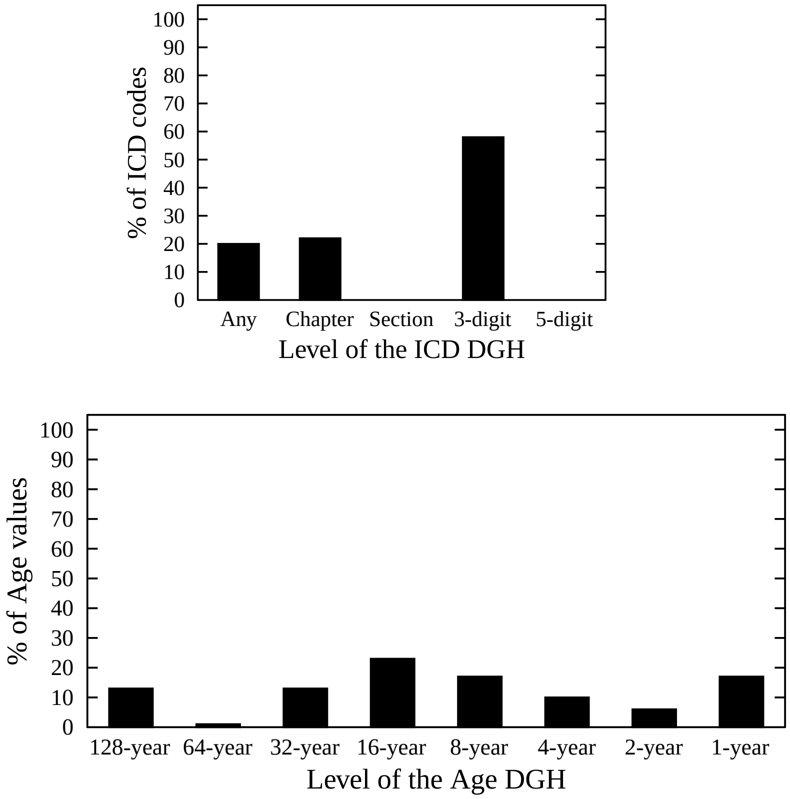


*Figure 14:* The histogram of ICD and Age after anonymizing $D_4^{Pop}$ using A-GS with $k = 5$.

first cluster convey more information than those in the second cluster. Specifically, we are certain that the patients in the first cluster were assigned the ICD code 250 when they were between 45 and 48 years old, later they were assigned an ICD code in the range 240 and 279 when they were 50 years old, and finally they were again assigned the ICD code 250 when they were between 57 and 64 years old. We are less certain about the profiles of the patients in the second cluster. We know that they were assigned an ICD code when they were 31 years old, assigned an ICD code in the range 240 and 279 when they were between 33 and 40 years old, assigned the ICD code 724 when they were 41 years old, and finally assigned an ICD code when they were between 49 and 52 years old.

## IV.5 Summary

In this chapter, we developed a framework to provide formal computational guarantees against re-identification attacks on longitudinal data. Our experiments verify that the proposed approach permits privacy-preserving publishing of longitudinal patient records while retaining the information of the original records signifi-

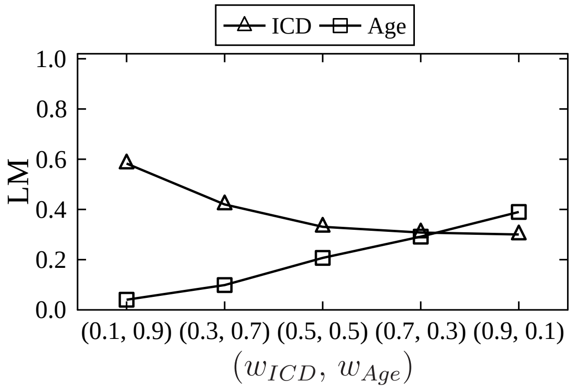


*Figure 15:* A comparison of information loss for $D_{50}^{Pop}$ using various $w_{ICD}$ and $w_{Age}$ values.

*Table 4:* Example clusters derived for $D_4^{Smp}$ using $k = 5$.

| |
|---|
| $(250,[45-48])$ $([240-279],50)$ $(250,[57-64])$ |
| $(250,[45-48])$ $([240-279],50)$ $(250,[57-64])$ |
| $(250,[45-48])$ $([240-279],50)$ $(250,[57-64])$ |
| $(250,[45-48])$ $([240-279],50)$ $(250,[57-64])$ |
| $(250,[45-48])$ $([240-279],50)$ $(250,[57-64])$ |

| |
|---|
| $([001-999],31)$ $([240-279],[33-40])$ $(724,41)$ $([001-999],[49-52])$ |
| $([001-999],31)$ $([240-279],[33-40])$ $(724,41)$ $([001-999],[49-52])$ |
| $([001-999],31)$ $([240-279],[33-40])$ $(724,41)$ $([001-999],[49-52])$ |
| $([001-999],31)$ $([240-279],[33-40])$ $(724,41)$ $([001-999],[49-52])$ |
| $([001-999],31)$ $([240-279],[33-40])$ $(724,41)$ $([001-999],[49-52])$ |

cantly more than a competitive baseline. This work is an important step towards increasing the type of information that can be made available to researchers while preserving patients' privacy. This is partly because our approach can be directly utilized by researchers when sharing longitudinal data for research with biorepositories.

# CHAPTER V

# DISCUSSION

In this chapter, we discuss how our methods can be extended to prevent a privacy threat in addition to re-identification. Additionally, we address the limitations of our work.

## V.1 Attacks Beyond Re-identification

Beyond re-identification is the threat of sensitive itemset disclosure, in which a patient is associated with a set of diagnosis codes, such as HIV, that reveal some sensitive information. Neither $k$-map nor $k$-anonymity guarantees preventing sensitive itemset disclosure, since a large number of records that are indistinguishable with respect to the potentially identifying diagnosis codes can still contain the same sensitive itemset [52]. We note that our methods can be extended to prevent this attack. This is possible by controlling generalization and suppression to ensure that an additional principle is satisfied, such as $\ell$-diversity [64], which dictates how sensitive information is grouped in the anonymized data. This extension, however, is beyond the scope of this thesis.

## V.2 Limitations

The work in this thesis is limited in certain aspects, which we highlight to suggest opportunities for further research. First, our algorithms do not limit the amount of information loss incurred by generalization and suppression. Designing algorithms that provide this type of guarantee is important to enhance the utility of anonymized datasets, but is also computationally challenging due to the large search spaces involved, particularly for longitudinal data. Second, our algorithms do not guarantee that the released data remain useful for scenarios in which pre-specified analytic tasks, such as the validation of known GWAS [23], are known to data owners a priori. To address such scenarios, we plan to modify our algorithms so that they take the tasks for which data are anonymized into account. We believe such tasks can be incorporated into our methods using *constraints* that denote which values are generalized together during anonymization. Third, we intend to evaluate the utility of the data our algorithms derive with other general and special information loss metrics, such as the ILoss metric [65] which is a variant of the LM and the classification metric [43] which measures the classification error on the anonymized data.

# CHAPTER VI

## CONCLUSIONS AND FUTURE WORK

The work in this thesis was motivated by the growing need to disseminate more detailed information to researchers without compromising patients' privacy rights. To the best of our knowledge, our methods for longitudinal data and the special case of repeated values are the first to allow sharing such data while providing computational privacy guarantees. Our investigations suggest that our methods can generate data that remain useful for various biomedical studies. This was illustrated through extensive experiments with real data derived from the EMRs of thousands of patients. The methods are not guided by specific utility (e.g., satisfaction of GWAS validation), but we are confident they can be extended to support such endeavors. We also note that further research is necessary to determine if our methods can derive data with an information loss that is sufficiently low for clinical phenotype discovery.

# BIBLIOGRAPHY

[1] D. Blumenthal, “Stimulating the adoption of health information technology,” *New England Journal of Medicine*, vol. 360, no. 15, pp. 1477–1479, 2009.

[2] D. A. Ludwick and J. Doucette, “Adopting electronic medical records in primary care: Lessons learned from health information systems implementation experience in seven countries,” *International Journal of Medical Informatics*, vol. 78, pp. 22–31, 2009.

[3] A. K. Jha, C. M. DesRoches, E. G. Campbell, K. Donelan, S. R. Rao, T. G. Ferris *et al.*, “Use of electronic health records in u.s. hospitals,” *New England Journal of Medicine*, vol. 360, pp. 1628–1638, 2009.

[4] B. B. Dean, J. Lam, J. L. Natoli, Q. Butler, D. Aguilar, and R. J. Nordyke, “Review: Use of electronic medical records for health outcomes research: A literature review,” *Medical Care Research and Review*, vol. 66, pp. 611–638, 2009.

[5] K. Holzer and W. Gall, “Utilizing ihe-based electronic health record systems for secondary use,” *Methods of Information in Medicine*, vol. 50, 2011, doi:10.3414/ME10-01-0060.

[6] C. Safran, M. Bloomrosen, W. Hammond, S. Labkoff, S. Markel-Fox, P. Tang *et al.*, “Toward a national framework for the secondary use of health data: an american medical informatics association white paper,” *Journal of the American Medical Informatics Association*, vol. 14, pp. 1–9, 2007.

[7] K. Tu, T. Mitiku, H. Guo, D. S. Lee, and J. V. Tu, “Myocardial infarction and the validation of physician billing and hospitalization data using electronic medical records,” *Chronic Diseases in Canada*, vol. 30, pp. 141–146, 2010.

[8] C. A. McCarty, R. L. Chisholm, C. G. Chute, I. J. Kullo, G. P. Jarvik, E. B. Larson *et al.*, “The emerge network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies,” *BMC Medical Genomics*, vol. 4, p. 13, 2011.

[9] I. J. Kullo, K. Ding, H. Jouni, C. Y. Smith, and C. G. Chute, “A genome-wide association study of red blood cell traits using the electronic medical record,” *Public of Library of Science ONE*, vol. 5, 2010.

[10] J. A. Pacheco, P. C. Avila, J. A. Thompson, M. Law, J. A. Quraishi, A. K. Greiman *et al.*, "A highly specific algorithm for identifying asthma cases and controls for genome-wide association studies," in *Proceedings of the 2009 American Medical Informatics Association Annual Symposium*, 2009, pp. 497–501.

[11] M. D. Ritchie, J. C. Denny, D. C. Crawford, A. H. Ramirez, J. B. Weiner, J. M. Pulley, M. A. Basford, K. Brown-Gentry, J. R. Balser, D. R. Masys, J. L. Haines, and D. M. Roden, "Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record," *American Journal of Human Genetics*, vol. 86, no. 4, pp. 560–572, 2010.

[12] M. N. Liebman, "Personalized medicine: a perspective on the patient, disease and causal diagnostics," *Personalized Medicine*, vol. 4, no. 2, pp. 171–174, 2007.

[13] N.-H. T. Trinh, S. J. Youn, J. Sousa, S. Regan, C. A. Bedoya, T. E. Chang, M. Fava, and A. Yeung, "Using electronic medical records to determine the diagnosis of clinical depression," *International Journal of Medical Informatics*, To be published, doi:10.1016/j.ijmedinf.2011.03.014.

[14] A. Tucker and D. Garway-Heath, "The pseudotemporal bootstrap for predicting glaucoma from cross-sectional visual field data," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 1, pp. 79–85, 2010.

[15] S. Jensen, "Mining medical data for predictive and sequential patterns," in *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2001, Discovery Challenge on Thrombosis.

[16] P. R. Burton, A. L. Hansell, I. Fortier, T. A. Manolio, M. J. Khoury, J. Little *et al.*, "Size matters: just how big is big?: Quantifying realistic sample size requirements for human genome epidemiology," *International Journal of Epidemiology*, vol. 38, pp. 263–273, 2009.

[17] C. C. Spencer, Z. Su, P. Donnelly, and J. Marchini, "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip," *Public of Library of Science Genetics*, vol. 5, no. 5, p. e1000477, 2009.

[18] National Institutes of Health, "Policy for sharing of data obtained in nih supported or conducted genome-wide association studies (gwas)," NOT-OD-07-088, 2007.

[19] M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov *et al.*, “The ncbi dbgap database of genotypes and phenotypes,” *Nature genetics*, vol. 39, no. 10, pp. 1181–1186, 2007.

[20] Department of Health and Human Services, “Standards for privacy of individually identifiable health information,” Final rule. Federal register; {45 CFR: Parts 160–164}, August 12, 2002.

[21] K. Benitez and B. Malin, “Evaluating re-identification risks with respect to the hipaa privacy rule,” *Journal of the American Medical Informatics Association*, vol. 17, no. 2, pp. 169–177, 2010.

[22] G. Loukides, J. C. Denny, and B. Malin, “The disclosure of diagnosis codes can breach research participants’ privacy,” *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 322–327, 2010.

[23] G. Loukides, A. Gkoulalas-Divanis, and B. Malin, “Anonymization of electronic medical records for validating genome-wide association studies,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 17, pp. 7898–7903, 2010.

[24] G. Loukides, A. Gkoulalas-Divanis, and B. Malin, “Privacy-preserving publication of diagnosis codes for effective biomedical analysis,” in *Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine*, 2010, pp. 1–6.

[25] A. Tamersoy, G. Loukides, J. C. Denny, and B. Malin, “Anonymization of administrative billing codes with repeated diagnoses through censoring,” in *Proceedings of the 2010 American Medical Informatics Association Annual Symposium*, 2010, pp. 782–786.

[26] L. Sweeney, “*k*-anonymity: A model for protecting privacy,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[27] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving data publishing: A survey of recent developments,” *ACM Computing Surveys*, vol. 42, no. 4, pp. 1–53, 2010.

[28] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala, “Privacy-preserving data publishing,” *Foundations and Trends in Databases*, vol. 2, no. 1-2, pp. 1–167, 2009.

[29] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, “Role-based access control models,” *IEEE Computer*, vol. 29, no. 2, pp. 38–47, 1996.

[30] B. Pinkas, “Cryptographic techniques for privacy-preserving data mining,” *ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations*, vol. 4, no. 2, pp. 12–19, 2002.

[31] S. M. Diesburg and A.-I. A. Wang, “A survey of confidential data storage and deletion methods,” *ACM Computing Surveys*, vol. 43, no. 1, pp. 1–37, 2010.

[32] N. R. Adam and J. C. Wortmann, “Security-control methods for statistical databases: A comparative study,” *ACM Computing Surveys*, vol. 21, no. 4, pp. 515–556, 1989.

[33] C. C. Aggarwal and P. S. Yu, “A survey of randomization methods for privacy-preserving data mining,” in *Privacy-Preserving Data Mining: Models and Algorithms*, ser. Advances in Database Systems. Springer, 2008, vol. 34, pp. 137–156.

[34] L. Willenborg and T. De Waal, “Statistical disclosure control in practice,” ser. Lecture Notes in Statistics. Springer, 1996, vol. 111, pp. 1–152.

[35] N. Marsden-Haug, V. Foster, P. Gould, E. Elbert, H. Wang, and J. Pavlin, “Code-based syndromic surveillance for influenzalike illness by international classification of diseases, ninth revision,” *Emerging Infectious Diseases*, vol. 13, pp. 207–216, 2007.

[36] K. E. Emam, “Data anonymization practices in clinical research: A descriptive study,” Access to Information and Privacy Division of Health in Canada, Tech. Rep., 2006.

[37] P. Samarati, “Protecting respondents’ identities in microdata release,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.

[38] K. El Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo *et al.*, “A globally optimal *k*-anonymity method for the de-identification of health data,” *Journal of the American Medical Informatics Association*, vol. 16, no. 5, pp. 670–682, 2009.

[39] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional *k*-anonymity,” in *Proceedings of the 22nd IEEE International Conference on Data Engineering*, 2006, pp. 25–35.

[40] T. Dalenius, “Finding a needle in a haystack – or identifying anonymous census record,” *Journal of Official Statistics*, vol. 2, no. 3, pp. 329–336, 1986.

[41] M. E. Nergiz and C. Clifton, "Thoughts on *k*-anonymization," *Data Knowledge and Engineering*, vol. 63, no. 3, pp. 622–645, 2007.

[42] G. Loukides and J. Shao, "Capturing data usefulness and privacy protection in *k*-anonymisation," in *Proceedings of the 22nd ACM Symposium on Applied Computing*, 2007, pp. 370–374.

[43] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 279–288.

[44] K. El Emam and F. K. Dankar, "Protecting privacy using *k*-anonymity," *Journal of the American Medical Informatics Association*, vol. 15, no. 5, pp. 627–637, 2008.

[45] B. Malin, "k-unlinkability: A privacy protection model for distributed data," *Data Knowledge and Engineering*, vol. 64, no. 1, pp. 294–311, 2008.

[46] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-preserving anonymization of set-valued data," in *Proceedings of the Very Large Data Bases Endowment*, vol. 1, no. 1, 2008, pp. 115–125.

[47] M. E. Nergiz, M. Atzori, Y. Saygin, and B. Guc, "Towards trajectory anonymization: a generalization-based approach," *Transactions on Data Privacy*, vol. 2, no. 1, pp. 47–75, 2009.

[48] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *Proceedings of the 24nd IEEE International Conference on Data Engineering*, 2008, pp. 376–385.

[49] M. Terrovitis and N. Mamoulis, "Privacy preservation in the publication of trajectories," in *Proceedings of the 9th International Conference on Mobile Data Management*, 2008, pp. 65–72.

[50] D. M. Roden, J. M. Pulley, M. A. Basford, G. R. Bernard, E. W. Clayton, J. R. Balser *et al.*, "Development of a large-scale de-identified dna biobank to enable personalized medicine," *Clinical Pharmacology and Therapeutics*, vol. 84, no. 3, pp. 362–369, 2008.

[51] G. Loukides and J. Shao, "Clustering-based *k*-anonymisation algorithms," in *Proceedings of the 18th International Conference on Database and Expert Systems Applications*, 2007, pp. 761–771.

[52] Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu, “Anonymizing transaction databases for publication,” in *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 767–775.

[53] M. Sullivan, *Fundamentals of Statistics*, 3rd ed. Pearson Prentice Hall, 2010.

[54] L. Sweeney, “Achieving *k*-anonymity privacy protection using generalization and suppression,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571–588, 2002.

[55] C. C. Aggarwal and P. S. Yu, “A condensation approach to privacy preserving data mining,” in *Proceedings of the 9th International Conference on Extending Database Technology*, 2004, pp. 183–199.

[56] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas *et al.*, “Achieving anonymity via clustering,” in *Proceedings of the 25th ACM Symposium on Principles of Database Systems*, 2006, pp. 153–162.

[57] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[58] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences,” *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[59] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. MIT Press, 2001.

[60] J. Domingo-Ferrer and J. M. Mateo-Sanz, “Practical data-oriented microaggregation for statistical disclosure control,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1, pp. 189–201, 2002.

[61] A. H. Ramirez, J. S. Schildcrout, D. L. Blakemore, D. R. Masys, J. M. Pulley, M. A. Basford *et al.*, “Modulators of normal electrocardiographic intervals identified in a large electronic medical record,” *Heart Rhythm*, To be published.

[62] J. C. Denny, M. D. Ritchie, D. C. Crawford, J. S. Schildcrout, A. H. Ramirez, J. M. Pulley, M. A. Basford, D. R. Masys, J. L. Haines, and D. M. Roden, “Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science,” *Circulation*, vol. 122, pp. 2016–2021, Nov 2010.

[63] C. C. Aggarwal, “On $k$-anonymity and the curse of dimensionality,” in *Proceedings of the 31st International Conference on Very Large Data Bases*, 2005, pp. 901–909.

[64] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “$\ell$-diversity: Privacy beyond $k$-anonymity,” in *Proceedings of the 22nd IEEE International Conference on Data Engineering*, 2006, pp. 24–35.

[65] X. Xiao and Y. Tao, “Personalized privacy preservation,” in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, 2006, pp. 229–240.