Enabling Health Data Sharing with Fine-Grained Privacy

Luca Bonomi*
Vanderbilt University Medical Center
Nashville, TN, USA
luca.bonomi@vumc.org

Sepand Gousheh
University of North Carolina at Charlotte
Charlotte, NC, USA
sgousheh@uncc.edu

Liyue Fan
University of North Carolina at Charlotte
Charlotte, NC, USA
liyue.fan@charlotte.edu

ABSTRACT

Sharing health data is vital in advancing medical research and transforming knowledge into clinical practice. Meanwhile, protecting the privacy of data contributors is of paramount importance. To that end, several privacy approaches have been proposed to protect individual data contributors in data sharing, including data anonymization and data synthesis techniques. These approaches have shown promising results in providing privacy protection at the dataset level. In this work, we study the privacy challenges in enabling fine-grained privacy in health data sharing. Our work is motivated by recent research findings, in which patients and healthcare providers may have different privacy preferences and policies that need to be addressed. Specifically, we propose a novel and effective privacy solution that enables data curators (e.g., healthcare providers) to protect sensitive data elements while preserving data usefulness. Our solution builds on randomized techniques to provide rigorous privacy protection for sensitive elements and leverages graphical models to mitigate privacy leakage due to dependent elements. To enhance the usefulness of the shared data, our randomized mechanism incorporates domain knowledge to preserve semantic similarity and adopts a block-structured design to minimize utility loss. Evaluations with real-world health data demonstrate the effectiveness of our approach and the usefulness of the shared data for health applications.

CCS CONCEPTS

• **Security and privacy** → **Data anonymization and sanitization**; • **Applied computing** → **Health informatics**; • **Information systems** → *Data mining*.

KEYWORDS

Data Sharing, Data Privacy, Health Data

ACM Reference Format:
Luca Bonomi, Sepand Gousheh, and Liyue Fan. 2023. Enabling Health Data Sharing with Fine-Grained Privacy. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3614864

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CIKM '23, October 21–25, 2023, Birmingham, United Kingdom
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614864

1 INTRODUCTION

The sharing of health data is vital in advancing medical research, enabling personalized medicine, and facilitating effective secondary data analysis [41]. However, the sensitive nature of health information poses significant privacy and ethical concerns in data sharing. Privacy regulations and policies (e.g., HIPAA) are put in place to protect the identity of data participants. As an example, the original data records may be stripped of the protected health information (PHI) via de-identification. Recently, there has been an increased need for regulations and technology to strengthen the privacy protection required by HIPAA [49]. For instance, without the constitutional protection for abortion access in the U.S., states may prosecute individuals (e.g., patients, healthcare providers) who seek or facilitate abortions. From a technological perspective, it is imperative to develop new privacy techniques that protect sensitive parts of the medical record.

Furthermore, providing fine-grained privacy protection empowers individuals with better privacy control, which could encourage data sharing [35]. For example, a recent study by Kim et al. [32] investigated patient privacy concerns and preferences toward data sharing. They found that more than 76% of the participants selected at least 1 condition to be protected (i.e., did not want to share), demonstrating the need of fine-grained privacy control. In general, honoring individual privacy preferences helps build trust and encourages data sharing behaviors in a broad range of domains [2, 45, 47].

In this work, we study the problem of fine-grained privacy control for health data sharing, while protecting elements (i.e., parts of a record) considered sensitive by the data curator and participants. Note that our privacy model is different from standard differential privacy [16], which aims to protect the presence of the entire record in the data. The new problem poses several privacy challenges. Firstly, the sole removal of sensitive elements in the data may not provide adequate privacy, as dependent elements that remain in the data may disclose information about the sensitive elements. Recent studies have shown that data dependence may lead to privacy breaches even with strong privacy models (e.g., differential privacy) [31, 37]. Addressing data dependence for health data sharing is an important issue, as health data often exhibit strong correlation (e.g., between medications, diagnoses, and procedures). Secondly, existing privacy techniques may result in overly-perturbed data, which can significantly reduce the usability of the shared data. Recent studies have shown that differential privacy may not only reduce data accuracy but also have undesired

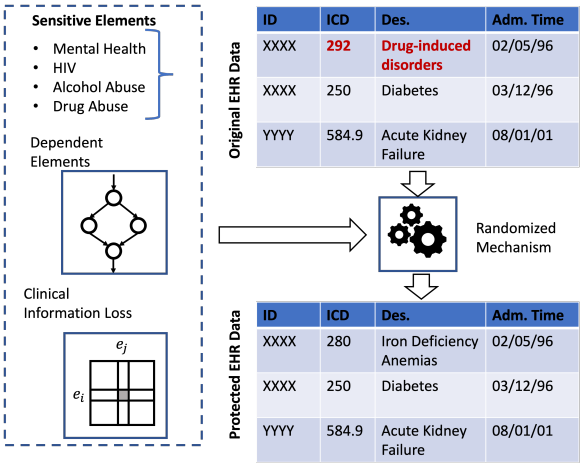


Figure 1: Illustration of the proposed solution. Our approach considers both dependent elements and the clinical information loss to design the randomized mechanism for privacy protection.

consequences on downstream applications (e.g., inflicting changes in the outcome [21, 55]).

We propose a new privacy-protecting approach for health data sharing, which provides rigorous protection for sensitive elements while retaining data usefulness. Specifically, our contributions are:

- We propose an effective approach to mitigate privacy breaches due to dependent elements. Specifically, we adopt probabilistic graphical models to capture data dependence and formulate an informed adversary, which infers sensitive elements from the observed data. Our method identifies non-sensitive elements that contribute most to the inference attack, and protects them together with the sensitive elements.
- We design a novel randomized mechanism to protect sensitive elements under the local differential privacy model. Our mechanism adopts a block structure and maps sensitive input and non-sensitive input differently. To retain the usefulness of health data, our approach incorporates the hierarchical representation of clinical concepts to report semantically similar data. We also propose an optimization problem formulation to find the best block structure.
- The empirical evaluation is conducted on real clinical data and practical sensitive conditions, which demonstrates the effectiveness of our proposed solution in comparison to state-of-the-art privacy practices. Furthermore, we conduct a case study for readmission risk prediction with the shared data, to illustrate the data usefulness for health applications.

The rest of the paper is organized as follows. Section 2 introduces local differential privacy and existing randomized response mechanisms; Section 3 describes our proposed approach to address privacy leakage due to dependent elements; Section 4 presents the proposed randomized response mechanism; Section 5 discusses evaluation results; Section 6 briefly reviews related work; Section 7 provides additional discussion around health data privacy and data sharing; Section 8 concludes the paper and discusses future work.

2 PRELIMINARIES

In this paper, we consider a patient's EHR record as a sequence of multiple clinical events (e.g., diagnosis, medications), ordered by the time in which they are recorded by the healthcare provider. Furthermore, we assume that the input domain of clinical events $\mathcal{X}$ can be divided into sensitive and non-sensitive elements: $\mathcal{X} = \mathcal{X}_S \bigcup \mathcal{X}_{NS}$. The set $\mathcal{X}_S$ contains commonly known sensitive clinical conditions such as mental health, alcohol abuse, and sexually transmitted diseases, which may be specified by patients and clinicians or defined by policies and regulations. The goal of fine-grained privacy protection is to share sanitized patient EHR records while protecting sensitive elements in $\mathcal{X}_S$.

2.1 Randomized Response for LDP

The problem studied in this work is closely related to the notion of local differential privacy (LDP) [15]. An obfuscation mechanism $M : \mathcal{X} \rightarrow \mathcal{Y}$ satisfies ϵ-LDP ($\epsilon \geq 0$) if for any $x, x' \in \mathcal{X}$ and any $y \in \mathcal{Y}$, if and only if the following inequality holds:

$$Pr[M(x) = y] \leq e^{\epsilon} \cdot Pr[M(x') = y]. \tag{1}$$

With $\mathcal{Y} = \mathcal{X}$, the $|\mathcal{X}|$-ary randomized response (RR) mechanism [30] has been shown to achieve ϵ-LDP with $Pr[RR(x) = y|x] = \frac{e^{\epsilon}}{|\mathcal{X}|+e^{\epsilon}-1}$ if $y = x$, and $\frac{1}{|\mathcal{X}|+e^{\epsilon}-1}$ otherwise. In a recent work, Murakami and Kawamoto [40] have proposed a utility maximizing randomized response mechanism (denoted as UMAXRR) to improve upon RR by providing ϵ-LDP only for sensitive data. Specifically, the input domain $\mathcal{X}$ is divided into sensitive elements $\mathcal{X}_S$ and non-sensitive elements $\mathcal{X}_{NS} = \mathcal{X} \setminus \mathcal{X}_S$. Similarly, the output range is divided into $\mathcal{Y}_P$, the protected set, and $\mathcal{Y}_I$, the invertible set. When $\mathcal{Y}_P = \mathcal{X}_S$ and $\mathcal{Y}_I = \mathcal{X}_{NS}$, the UMAXRR mechanism provides ϵ-LDP for input elements in $\mathcal{X}_S$. The mechanism is defined as:

$$Pr[y|x] = \begin{cases} \frac{e^{\epsilon}}{|\mathcal{X}_S|+e^{\epsilon}-1} & \text{if } x \in \mathcal{X}_S \text{ and } y = x, \\ \frac{1}{|\mathcal{X}_S|+e^{\epsilon}-1} & \text{if } x \in \mathcal{X}_S \text{ and } y \in \mathcal{Y}_P \setminus \{x\}, \\ \frac{1}{|\mathcal{X}_S|+e^{\epsilon}-1} & \text{if } x \in \mathcal{X}_{NS} \text{ and } y \in \mathcal{Y}_P, \\ \frac{e^{\epsilon}-1}{|\mathcal{X}_S|+e^{\epsilon}-1} & \text{if } x \in \mathcal{X}_{NS} \text{ and } y = x, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

It is important to point out that performing RR on $\mathcal{X}_S$ may only provide ϵ-indistinguishability among sensitive elements, but it does not prevent an adversary from learning the occurrence of sensitive elements in the data. In contrast, UMAXRR ensures that both sensitive input and non-sensitive input can map to sensitive output in $\mathcal{Y}_P$. By observing an output in $\mathcal{Y}_P$, the adversary's inference of whether the input is sensitive vs. non-sensitive is also bounded (a similar principle has been discussed in one-sided differential privacy [33]) We will show similar guarantees for our proposed mechanism in Theorem 1. Murakami and Kawamoto [40] have shown that UMAXRR outperforms RR in frequency estimation tasks, due to relaxed privacy protection. It can be seen in Equation 2 that whenever an adversary observes an output in $\mathcal{Y}_I$ (recall that $\mathcal{Y}_I = \mathcal{X}_{NS}$), the input can be inferred with certainty, i.e., $x = y$.

2.2 Challenges and Solution Overview

Although UMAXRR may provide utility improvement over RR, there exist significant challenges in its application in practice. Firstly,

any non-sensitive data reported by the UMAXRR mechanism is truthful. While this benefits the data usability, it may also lead to privacy breaches for sensitive elements due to dependence in longitudinal data. In fact, there may be temporal correlations between non-sensitive data and sensitive data in real applications (e.g., mobile applications [24]). As a result, an adversary may infer a user's sensitive data by leveraging the observed non-sensitive data and their correlation. Secondly, UMAXRR may inflict high utility loss on sensitive data. In fact, when the sensitive domain $|\mathcal{X}_S|$ is large, the probability of preserving the original data is low, as can be seen in Equation 2. Furthermore, the mechanism has equal chance of reporting any $y \in \mathcal{Y}_P \setminus \{x\}$, without considering the semantic similarity among data elements.

Our proposed approach, depicted in Figure 1, presents two main contributions in addressing aforementioned challenges. To bridge the gap on protecting sensitive elements in longitudinal data, our approach leverages graphical models for data dependence and efficiently extends LDP protection to selected non-sensitive elements. To improve data utility for health applications, we design a novel randomized response approach that leverages the semantic similarity in clinical data domain and minimizes utility loss. In the following, we present how our solution addresses dependent data and describe the proposed randomized response mechanism.

3 MITIGATING PRIVACY INFERENCE DUE TO DEPENDENT ELEMENTS

As an adversary may leverage data dependence to infer or reconstruct sensitive elements, obfuscating sensitive elements alone may not provide sufficient privacy protection. As an example, the presence of a non-sensitive medication in the shared data (e.g., insulin) may enable the adversary to infer a sensitive condition that has been obfuscated (e.g., diabetes). In this section, we describe how to quantify the privacy risk due to dependent elements and present an effective solution.

Modeling Data Dependence. To model the dependence between elements in health data, we consider a graphical model $G = (V, E)$ (e.g., Bayesian network), where each node $e \in V$ represents a distinct element (e.g., diagnosis code, medication), and a directed edge $(e_i, e_j) \in E$ from e_i to e_j represents that e_j is observed after e_j in a patient record within ΔT time units. Furthermore, each edge (e_i, e_j) has a weight $\delta(e_i, e_j)$, which captures the conditional probability $Pr[e_j|e_i]$ between elements e_i and e_j. In practice, the graphical model can be constructed with auxiliary information, such as using previous studies or publicly available datasets. In our study, we reserve a portion of the overall data to learn the graphical model, which is disjoint from the data used for evaluations.

Adversarial Model. We consider an informed adversary who has knowledge of the universe of sensitive elements (i.e., $\mathcal{X}_S$), the privacy mechanism for data obfuscation (e.g., randomized response), and the dependence in the graphical model G. Given an obfuscated element $\hat{e}$ and the set of non-sensitive elements in the user's output data O_{NS} (where $O_{NS} \subseteq \mathcal{Y}_I = \mathcal{X}_{NS}$), the goal of the adversary is to infer the original value of the sensitive element e. Formally, the adversary can compute the posterior probability of $e = e_i$, as follows:

$$Pr[e = e_i|\hat{e}, O_{NS}, G] = \frac{Pr[\hat{e}, G, O_{NS}|e = e_i]Pr[e = e_i]}{Pr[\hat{e}, G, O_{NS}]} = \frac{Pr[\hat{e}|G, O_{NS}, e = e_i]}{Pr[\hat{e}|G, O_{NS}]} Pr[e = e_i|G, O_{NS}].$$

Hence, the adversary can maximize the posterior belief with:

$$e^* = arg \max_{e_i \in \mathcal{X}_S} \{ \frac{Pr[\hat{e}|G, O_{NS}, e = e_i]}{Pr[\hat{e}|G, O_{NS}]} Pr[e = e_i|G, O_{NS}] \}. \quad (3)$$

In the equation above, the term $Pr[\hat{e}|G, O_{NS}, e = e_i]$ captures the information revealed by the obfuscation mechanism (i.e., the probability of reporting $\hat{e}$ given G and input e_i). Similarly, $Pr[\hat{e}|G, O_{NS}]$ represents the probability of reporting $\hat{e}$ given G and any input. The term $Pr[e = e_i|G, O_{NS}]$ captures the data dependence modeled by the graphical model G. In the next section, we will show that our privacy mechanism provides LDP guarantees for sensitive elements, which ensures that the ratio $\frac{Pr[\hat{e}|G,O_{NS},e=e_i]}{Pr[\hat{e}|G,O_{NS}]}$ is bounded. Therefore, we must take into consideration of $Pr[e = e_i|G, O_{NS}]$ to mitigate the adversarial inference in Equation 3.

Mitigating Privacy Leakage. To reduce the privacy leakage due to data dependence, one simple solution is to extend the LDP protection to any element in $\mathcal{X}$ such that there is a path (either forward or backward) in G connecting it to a sensitive element in $\mathcal{X}_S$. In other words, we will expand the sensitive domain $\mathcal{X}_S$, to include both the initial sensitive elements and additional elements that have a path to them. While this simple solution effectively limits the privacy inference, it may lead to overly-perturbed data if a large number of elements are added to the sensitive domain.

To mitigate the privacy inference risk while preserving non-sensitive elements, we propose to adopt a threshold γ_i for each sensitive element $e_i \in \mathcal{X}_S$ and determine whether some non-sensitive elements should be also protected. Our goal is to upper-bound the term $Pr[e = e_i|G, O_{NS}]$ in Equation 3. As $O_{NS} \subseteq \mathcal{X}_{NS}$, we argue that it is sufficient to bound $Pr[e = e_i|G, \mathcal{X}_{NS}] \leq \gamma_i$. When sensitive elements have strong correlation with certain non-sensitive elements, as modeled by G, this bound may not hold. Therefore, we propose to select non-sensitive data elements to remove from $\mathcal{X}_{NS}$ (i.e., adding to $\mathcal{X}_S$), such that the observations of those elements in the output data may no longer be truthful, and therefore cannot be used by the adversary to reliably launch inference attacks.

Previous works have shown that finding a minimum set of non-sensitive data to hide in order to protect sensitive data is a challenging problem [1, 8, 51]. A simple brute-force approach would examine all possible subsets of non-sensitive data, which may be exponential with $|\mathcal{X}|$. In this work, we estimate $Pr[e = e_i|G, \mathcal{X}_{NS}] \approx \sum_{e' \in \mathcal{X}_{NS}} Pr[e_i|e']Pr[e']$, where $Pr[e_i|e']$ is the conditional probability which accounts for all paths from e' to e_i, and $Pr[e']$ is estimated with the empirical frequency. We adopt a greedy approach to iteratively select non-sensitive elements that contribute most to the privacy leakage, i.e., highest $\sum_{e_i \in \mathcal{X}_S} Pr[e_i|e']Pr[e']$, and remove them from $\mathcal{X}_{NS}$, until the privacy leakage $Pr[e = e_i|G, \mathcal{X}_{NS}]$ is bounded by γ_i, $\forall e_i \in \mathcal{X}_S$.

4 RANDOMIZED OBFUSCATION TO ENHANCE DATA USEFULNESS

In this section, we describe the proposed randomized obfuscation mechanism M that reports sensitive elements $\mathcal{X}_S$ and non-sensitive $\mathcal{X}_{NS}$ differently. We adopt the same partition of the output range as in [40] with $\mathcal{Y}_P = \mathcal{X}_S$ and $\mathcal{Y}_I = \mathcal{X}_{NS}$. Note that $\mathcal{X}_S$ (hence $\mathcal{Y}_P$) may contain additional elements, which have been added as a result of bounding privacy leakage in Section 3 or adaptive blocking described later in this section.

The key innovation of our approach is that it takes advantages of the hierarchical representation of clinical data to quantify the semantic similarity and calibrates the randomization process to improve utility. As a result, our approach generates obfuscated data that are semantically similar to real data. We will show that our approach ensures LDP protection for elements in $\mathcal{X}_S$, similar to the guarantees in [40].

4.1 Measuring Clinical Information Loss

In EHR data, clinical concepts (e.g., diagnosis codes) are structured into taxonomy trees, in which nodes within the same subtree are similar to each other. For example, the International Classification of Diseases (ICD) introduces a tree-based model to group diagnosis codes into groups representing related diagnoses (e.g., diseases of the circulatory system, diseases of the respiratory system, etc.). With a tree-based representation, researchers have proposed several concept-level similarity measures [13, 27], such as using the lowest common ancestor for a pair of nodes in the taxonomy tree to capture their dissimilarity. Building on those techniques, we propose the following dissimilarity measure between data elements. Given e_i and e_j (e.g., ICD diagnosis codes), their dissimilarity is defined by

$$d(e_i, e_j) = \begin{cases} 1 & \text{if } e_i \text{ or } e_j \text{ is root,} \\ \alpha \frac{|\mathcal{L}(c)|-1}{n-1} + (1-\alpha)\frac{|idx_c(e_i) - idx_c(e_j)|}{|\mathcal{L}(c)|} & \text{otherwise.} \end{cases} \tag{4}$$

where $c = LCA(e_i, e_j)$ denotes the lowest common ancestor for e_i and e_j, $\mathcal{L}(c)$ denotes the set of leaves of the subtree rooted by c, $idx_c(e)$ represents the relative order of e in a sorted $\mathcal{L}(c)$, and n denotes the total number of leaves in the tree. The proposed dissimilarity measure in Equation 4 is bounded between 0 and 1, where elements with lower dissimilarity values represent similar concepts. As an example, Figure 2 depicts the pair-wise dissimilarity for ICD9 codes obtained with the proposed measure.

We adopt this dissimilarity measure to quantify the utility loss when obfuscating e with $\hat{e}$, which distinguishes our setting to that of RR and UMAXRR where utility is lost completely if $\hat{e} \neq e$. To that end, we define the expected clinical information loss (CIL) for any $e \in \mathcal{X}_S$ inflicted by an obfuscation method M as follows:

$$\mathbb{E}[CIL(e)] = \sum_{\hat{e} \in \mathcal{Y}_P} d(e, \hat{e}) Pr[M(e) = \hat{e}] \tag{5}$$

which will be used in the next subsection to improve the utility of the privacy mechanism.

4.2 Randomized Obfuscation Mechanism

In the following, we first present the proposed block-structured mechanisms for elements in the sensitive set and the non-sensitive set. Then, we provide the privacy guarantees of the proposed mechanisms. Lastly, we describe an optimization problem formulation, which identifies the block structure that minimizes the utility loss as in Equation 5.

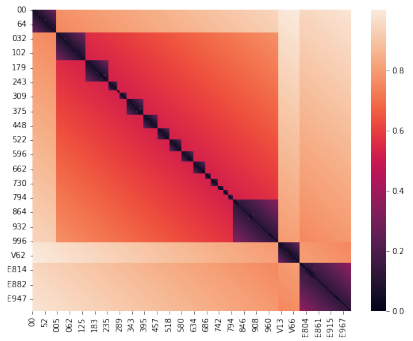


Figure 2: Visualization of the semantic similarity measure for ICD9 codes. The presented dissimilarity matrix has a block structure, which represents codes under various categories. Smaller values indicate stronger clinical semantic similarity.

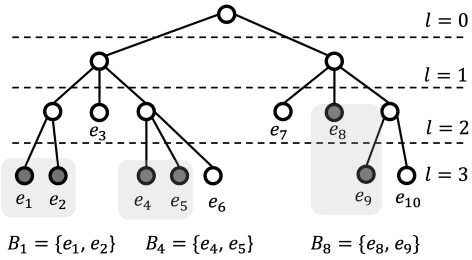


Figure 3: An illustrate example of blocking with size 2 and $\mathcal{X}_S = \{e_1, e_2, e_4, e_5, e_8, e_9\}$. For each sensitive element, a block containing most similar elements is created by traversing the taxonomy tree to compute the dissimilarity measure in Eq. 4. For example, for e_1, its block B_1 contains e_1 and e_2, while for e_8, its block B_8 contains e_8 and e_9. Intuitively, the randomized mechanism aims at sampling within the same block to retain semantic similarity.

Mechanism for Sensitive Elements. Our idea is to define blocks in the output domain with semantically similar elements and then map a sensitive element according to its block. An illustrative example for blocking is reported in Figure 3.

Given block size $b > 1$, we can construct a block B_i for every $e_i \in \mathcal{X}_S$ such that $B_i \subset \mathcal{Y}_P$ and B_i contains e_i as well as $b-1$ elements in $\mathcal{Y}_P$ most similar to e_i. We proposed the following randomized mechanism M when the input is e_i:

$$Pr[M(e_i) = \hat{e}] = \begin{cases} \frac{p_T}{b} & \text{if } \hat{e} \in B_i, \\ \frac{1-p_T}{|\mathcal{Y}_P|-b} & \text{if } \hat{e} \in \mathcal{Y}_P \setminus B_i, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Note that $p_T \in [0, 1]$ is dependent on the privacy parameter ϵ, which we will discuss in Theorem 1. The block-structured mechanism can be considered as a generalization of the plain randomized response, in which the input sensitive element is mapped within their block or to any other elements outside the block. Figure 3 provides an example for blocking. In the example, the sensitive

element e_1 can be mapped to similar elements in B_1 with probability p_T, while it can be mapped to other sensitive elements with probability $1 - p_T$. Intuitively, the block-structured mechanism can reduce the expected clinical information loss on the original sensitive element, as the obfuscated element is more likely to be sampled among semantically similar ones (with a calibrated p_T).

Mechanism for Non-Sensitive Elements. For a non-sensitive element $e_i \in \mathcal{X}_{NS}$, we propose the following randomized response, which reports a sensitive element at random or reports the input element. Note that $p_S \in [0, 1]$ will be discussed in Theorem 1.

$$Pr[M(e_i) = \hat{e}] = \begin{cases} \frac{p_S}{|\mathcal{Y}_P|} & \text{if } \hat{e} \in \mathcal{Y}_P, \\ (1 - p_S) & \text{if } \hat{e} = e_i, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Our approach satisfies ϵ-LDP for sensitive elements in $\mathcal{X}_S$. Specifically, the following Theorem provides the privacy analysis for our proposed mechanisms. Algorithm 1 reports the procedure of the combined mechanism.

Theorem 1. *The proposed mechanisms in Equation 6 and 7 provide the following ϵ-LDP guarantee:*

$$\frac{Pr[M(e_i) = \hat{e}]}{Pr[M(e_j) = \hat{e}]} \leq e^{\epsilon}, \forall e_i, e_j \in \mathcal{X}, \hat{e} \in \mathcal{Y}_P$$

when $p_T \leq \frac{e^{\epsilon} b}{|\mathcal{Y}_P| - b + e^{\epsilon} b}$ *and* $p_S \geq \frac{|\mathcal{Y}_P|}{e^{\epsilon}} \max\left\{\frac{p_T}{b}, \frac{1-p_T}{|\mathcal{Y}_P|-b}\right\}$.

Proof. We consider the following cases.

- Case 1: $e_i, e_j \in \mathcal{X}_S$, $\hat{e} \in B_i$ and $\hat{e} \notin B_j$.
$$\frac{Pr[M(e_i) = \hat{e}]}{Pr[M(e_j) = \hat{e}]} = \frac{p_T/b}{(1 - p_T)/(|\mathcal{Y}_P| - b)}$$
- Case 2: $e_i, e_j \in \mathcal{X}_S$, $\hat{e} \notin B_i$ and $\hat{e} \notin B_j$.
$$\frac{Pr[M(e_i) = \hat{e}]}{Pr[M(e_j) = \hat{e}]} = \frac{1 - p_T}{1 - p_T} \frac{|\mathcal{Y}_P| - b}{|\mathcal{Y}_P| - b} = 1$$
- Case 3: $e_i, e_j \in \mathcal{X}_S$ with $e_i, e_j \neq \hat{e}$, $\hat{e} \in B_i$ and $\hat{e} \in B_j$
$$\frac{Pr[M(e_i) = \hat{e}]}{Pr[M(e_j) = \hat{e}]} = \frac{p_T}{p_T} \frac{b}{b} = 1$$
- Case 4: $e_i \in \mathcal{X}_S$ and $e_j \in \mathcal{X}_{NS}$, $\hat{e} \in B_i$.
$$\frac{Pr[M(e_i) = \hat{e}]}{Pr[M(e_j) = \hat{e}]} = \frac{p_T/b}{p_S/|\mathcal{Y}_P|}$$
- Case 5: $e_i \in \mathcal{X}_S$ and $e_j \in \mathcal{X}_{NS}$, $\hat{e} \notin B_i$.
$$\frac{Pr[M(e_i) = \hat{e}]}{Pr[M(e_j) = \hat{e}]} = \frac{(1 - p_T)/(|\mathcal{Y}_P| - b)}{p_S/|\mathcal{Y}_P|}$$

To ensure ϵ-LDP, we need to ensure that the ratio in every case above is upper bounded by e^{ϵ}. Specifically, from case 1 we have:

$$p_T \leq \frac{e^{\epsilon} b}{|\mathcal{Y}_P| - b + e^{\epsilon} b}$$

From case 4 and case 5, we have:

$$p_S \geq \frac{|\mathcal{Y}_P|}{e^{\epsilon}} \max\left\{\frac{p_T}{b}, \frac{1 - p_T}{|\mathcal{Y}_P| - b}\right\}$$

Thus, by setting p_T and p_S as above, the ϵ-LDP protection as defined holds for all values of $\epsilon \geq 0$. □

Adaptive Blocking to Minimize Utility Loss. Here we discuss how to determine the best block structure, i.e., size b and $\mathcal{B} = \{B_i\}$, in order to minimize the utility loss of the randomized mechanism M. Given the privacy parameter ϵ, we are interested in finding the blocking $\mathcal{B}$ for M that minimizes the clinical information loss while satisfying ϵ-LDP for sensitive elements. This is formulated as the following optimization problem:

$$\begin{aligned} \min \quad & \sum_{e \in \mathcal{X}_S} \mathbb{E}[CIL(e)] \text{ as in Eq. 5} \\ \text{s.t.} \quad & Pr[M_{\mathcal{B}}(e_i) = \hat{e}] \leq e^{\epsilon} Pr[M_{\mathcal{B}}(e_j) = \hat{e}] \quad \forall e_i, e_j \in \mathcal{X}, \hat{e} \in \mathcal{Y}_P \end{aligned} \tag{8}$$

where we use $M_{\mathcal{B}}$ explicitly to denote the instantiation of the randomized mechanism with the block set $\mathcal{B}$. Intuitively, as the block size b increases, the likelihood of reporting the input sensitive element truthfully tends to decrease. However, if each block B_i contains elements that are sufficiently similar to the sensitive element e_i, the expected clinical usability loss may be reduced. Moreover, it is likely that the elements in $\mathcal{X}_S$ may not be similar to each other, e.g., ICD9 codes under distinct categories. Therefore, we hypothesize that it may be beneficial to further expand $\mathcal{X}_S$ by including selected non-sensitive elements to minimize the utility loss.

We propose an heuristic approach for the optimization problem in Equation 8. Given ϵ, our approach starts with $\mathcal{X}_S$ and $b = 1$, i.e., $B_i = \{e_i\}, \forall e_i \in \mathcal{X}_S$, and progressively increases the block size b and extends the set of sensitive elements, with the goal of finding the configuration that minimizes the objective in Equation 8. Specifically, we iteratively increase b with the current $\mathcal{X}_S$ and update the block set $\mathcal{B}$, until increasing b alone will no longer reduce the expected CIL. At that point, we select one non-sensitive element with the smallest average dissimilarity to elements in $\mathcal{X}_S$, add it to $\mathcal{X}_S$, and update the block set $\mathcal{B}$. The search process terminates when all the blocking configurations up to a maximum block size $b_{max} (\leq |\mathcal{Y}_P|)$ are evaluated, returning the configuration with minimum expected CIL. Note that the best blocking $\mathcal{B}$ highly depends on the privacy parameter ϵ, e.g., $B_i = \{e_i\}$ may be optimal for larger ϵ values, as shown in the evaluations in the next section.

To summarize, we first extend the sensitive set $\mathcal{X}_S$ to mitigate privacy leakage due to dependent element (described in Section 3); next, we identify the blocking structure $\mathcal{B}$ and possibly further extend $\mathcal{X}_S$ to minimize the expected clinical information loss. Neither requires access to private data. In fact, the graphical model G and priors can be estimated with auxiliary information disjoint from private data, inflicting no additional privacy cost.

5 RESULTS

Our evaluation centers around the practical needs for protecting sensitive clinical data. Specifically, we note that Title 38 U.S.C. § 7332 requires the United States Department of Veterans Affairs (VA) to obtain a signed authorization from Veterans whose health record contains sensitive conditions, including drug abuse, alcoholism and alcohol abuse, human immunodeficiency virus (HIV) infection, and sickle cell anemia, prior sharing their data [9, 14]. Hence, we initialize $\mathcal{X}_S$ with 45 ICD9 codes representing the sensitive conditions. We assess the effectiveness of several obfuscation methods in protecting these sensitive conditions.

Algorithm 1 RandomizedObfuscation($\mathcal{X}_S, \epsilon, e, \mathcal{B}$)

Require: $\mathcal{X}_S$ - set of sensitive elements, ϵ - privacy parameter, e - input element, $\mathcal{B} = \{B_1, B_2, \dots\}$ - blocks defined for all sensitive elements.

$\mathcal{Y}_P \leftarrow \mathcal{X}_S$
$b \leftarrow$ blocksize
$p_T \leftarrow \frac{e^\epsilon b}{|\mathcal{Y}_P| - b + e^\epsilon b}$
$p_S \leftarrow \frac{|\mathcal{Y}_P|}{e^\epsilon} \max\left\{\frac{p_T}{b}, \frac{1-p_T}{|\mathcal{Y}_P|-b}\right\}$
if $e \in \mathcal{X}_S$ **then** ▷ Sensitive element
    identify the block B defined for e
    sample output $\hat{e}$ according to Eq. 6
else ▷ Non-sensitive element
    sample output $\hat{e}$ according to Eq. 7
end if
return $\hat{e}$ ▷ Obfuscated element

Data. We report the evaluation on a de-identified real-word clinical dataset MIMIC-III [28], which comprises over 58,000 ICU hospital admissions for 38,645 adults and 7,875 neonates. As the domain of clinical data is large, we adopt the common practice of considering only the first three digits of the ICD9 codes. For each patient, we construct a sequence of temporal events obtained by selecting the primary and secondary diagnosis codes reported at each admission. Among all the patients, we focus on those with at least two admissions. Data are further partitioned with 80% for learning G and 20% for evaluating the privacy mechanisms. ΔT is set to 300 days to capture long-term correlation.

Obfuscation Approaches. We consider three different obfuscation approaches. As a baseline, we consider suppression, in which all sensitive elements are replaced by a special symbol (i.e., the root of the ICD hierarchical representation) while all non-sensitive elements are disclosed. Second, we consider the utility maximizing randomized response (UMAXRR) proposed in [40], where both sensitive and non-sensitive elements are randomized. Similar to our solution, this approach provides ϵ-LDP guarantee only for sensitive elements. Finally, we assess our proposed approach, named ClinSimRR, in which the obfuscation mechanism leverages both the dependence between data elements and their clinical similarity.

Parameters. The parameter ϵ controls the privacy protection provided by the randomized-based algorithms, where smaller values indicate stronger privacy. In our settings, we consider $\epsilon \in [0.1, 5.0]$. The parameter γ impacts the number of non-sensitive elements that need to protected to mitigate privacy leakages due to dependent elements. Lower values of γ lead to more elements to be obfuscated by our algorithm. In our evaluations, we consider $\gamma \in \{0.01, 0.05, 0.1\}$.

5.1 Metrics

We consider a variety of metrics to assess the usefulness and privacy protection of the shared data. Each experiment is conducted over 25 runs, and the results are reported with a 95% confidence interval.

Clinical Information Loss (CIL). To quantify the clinical usefulness of the shared admission data, we measure the average clinical information loss with respect to the original admissions. Specifically, we use the definition of dissimilarity $d(e_i, e_j)$ from Section 4.1 to quantify the clinical information loss. Let $R_j = (e_1, e_2, \dots, e_n)$ and $\hat{R}_j = (\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n)$ denote the original and sanitized records for the j-th patient, respectively. Then, we compute the patient-level clinical information loss as $\frac{1}{n}\sum_{i=1}^{n} d(e_i, \hat{e}_i)$. Lower values indicate that the clinical usefulness is well retained in the shared data. We report the measure for patients who exhibit at least one sensitive element in the original data.

Kullback–Leibler (KL) Divergence. To measure the usefulness at dataset level, we use the Kullback–Leibler (KL) divergence, which quantifies the dissimilarity between the probability distributions for the elements in the original and obfuscated data. Lower KL divergence values indicate a higher similarity in element distributions between the original and sanitized data.

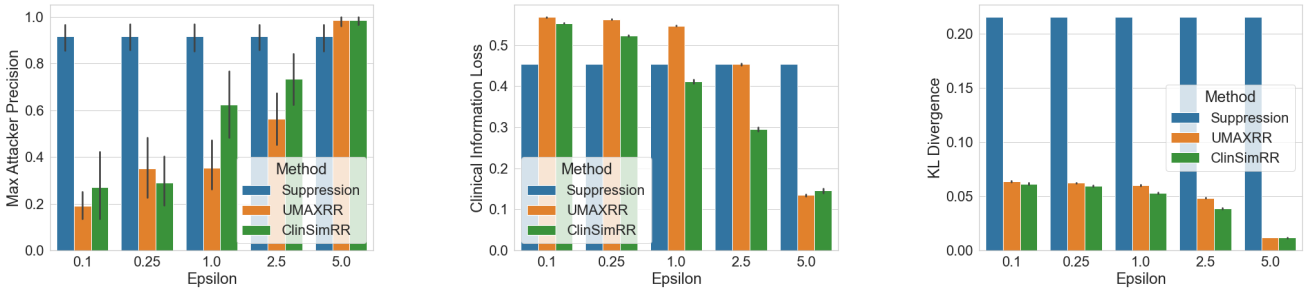
Maximum Attacker's Precision (MAP). We conduct an empirical privacy evaluation to quantify the privacy risk under dependent elements. Specifically, we consider an adversary who infers the unknown sensitive values using the inference attack model described in Eq.(3). We report the maximum precision (MAP) over the sensitive elements, where higher values indicate greater success in correctly reconstructing the obfuscated sensitive elements.

5.2 Comparing Obfuscation Approaches

Impact of the Privacy Parameters. Figure 4 reports a comparative evaluation by considering different values of the privacy parameter ϵ. Our results in Figure 4a show that suppression incurs higher privacy risks compared to randomized methods, where in the worst case the attacker can successfully infer a suppressed sensitive element from the shared data with precision above 0.8. Both utility maximizing randomized response and our approach significantly reduce the maximum attacker's precision, as the privacy protection increases (i.e., lower values of ϵ parameter). From Figure 4b, we observe that our approach provides better utility, i.e., lower clinical information loss, compared to UMAXRR. When privacy is relaxed (i.e., higher values of ϵ), both randomized approaches incur lower clinical information loss compared to suppression, as the original sensitive values are more likely to be retained. Finally, we observe that the randomized approaches better preserve the data distribution compared to suppression, leading to lower values of KL divergence in Figure 4c.

Figure 5a reports the privacy risk in terms of maximum attacker's precision by varying the parameter γ. With smaller values of γ, our approach extends the privacy protection to more dependent elements in the shared data. As a result, our mechanism may operate on a larger domain of protected elements compared to suppression and UMAXRR. While the extended set of protected elements can help mitigate privacy breaches, the usability of the shared data may decrease (i.e., higher CIL), as illustrated in Figure 5b. Overall, the parameter γ can be tuned to reduce privacy leakage due to dependent elements. We notice that γ has a larger impact in reducing the privacy risks in high privacy regimes (i.e., small values of ϵ). This is because for larger values of ϵ, sensitive input elements are more likely to be retained by the mechanism, thus the inference attack may rely less on the dependent elements.

Impact of the Number of Sensitive Elements. We study whether the number of sensitive elements impacts the usefulness and privacy of the shared data. Specifically, we vary the number of sensitive codes to be protected from the original set of 45 codes. Figure 6a

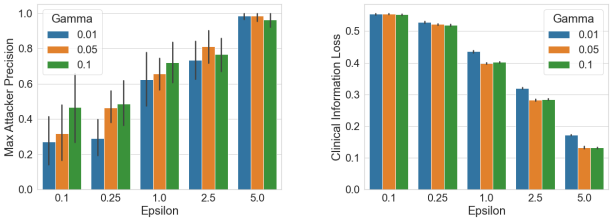


(a) Privacy risk of the shared data measured in terms of maximum attacker's precision.

(b) Usability of the shared data measured in terms of clinical information loss.

(c) Usability of the shared data measured in terms of KL-divergence.

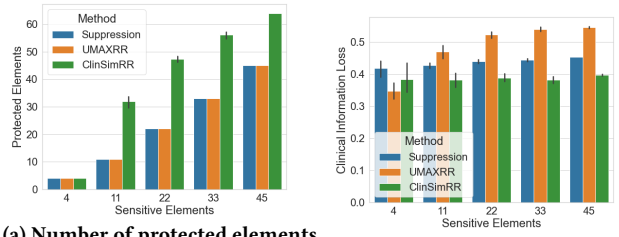
Figure 4: Impact of the privacy parameter (ϵ) on the utility and privacy of the shared data generated by suppression, utility maximizing randomized response, and our solution.



(a) Attacker's precision.

(b) Clinical information loss.

Figure 5: Impact of γ on the privacy risk and usability of the shared data.



(a) Number of protected elements adopted by obfuscation methods.

(b) Utility of shared data.

Figure 6: Impact of the number of sensitive elements. $\epsilon = 1.0$ for UMAXRR and our solution.

reports the overall number of protected elements adopted by the privacy approaches, with an increasing number of sensitive elements chosen from the original list of sensitive codes in input. Suppression and UMAXRR do not modify the input sensitive set. However, our approach may extend the privacy protection to a greater number of elements to mitigate privacy breach due to dependent elements and to minimize utility loss. The extended number of protected elements in our approach may vary as illustrated in Figure 6a, due to data dependence, the initial sensitive set, and the privacy parameters. Despite protecting a larger number of elements, our mechanism reports highly useful data especially when the sensitive set is large, illustrated by lower clinical information loss in Figure 6b.

Adaptive Blocking Strategy. We evaluate the effectiveness of our adaptive blocking strategy with different values of ϵ. Figure 7 shows an histogram representing the average clinical information loss for all the protected elements (extended $\mathcal{X}_S$) and a line graph representing the loss only for the 45 sensitive conditions. As the privacy protection is relaxed, we observe that the clinical information loss decreases for both sets. Additionally, the line graph reports the block size (blue label) adaptively selected by our approach, as the privacy parameter ϵ varies. Specifically, we observe that for low privacy regimes (i.e., large values of ϵ) our algorithm relies on small blocks, while in high privacy regimes (i.e., small values of ϵ) larger blocks are constructed. By using an adaptive blocking strategy, the set of protected elements may be extended to include additional elements to create better blocking. Overall, thanks to the adaptive blocking strategy, our method finds the best blocking structure to improve the usability of the shared data.

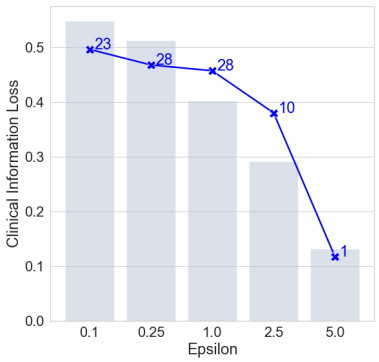


Figure 7: Clinical information loss (CIL) by our approach with different ϵ values for extended protected elements (histogram) and 45 sensitive conditions (solid line). Text label indicates the block size calibrated by our adaptive blocking strategy for each ϵ.

5.3 Case Study: Readmission Risk Analysis

Obfuscation methods considered in this work protect patient privacy by obscuring sensitive elements in EHR data. However, those sensitive elements may carry critical information for downstream clinical decision support applications. As an example, analysis of the shared data may enable a timely and effective identification of

patients at risk of readmission, and can benefit patient care, improve health outcomes, and reduce costs. It is important to study the usability of obfuscation methods in predicting patients at risk of readmission. In this case study, we adopt the same 45 sensitive conditions as described at the beginning of this section, and investigate the potential utility loss in readmission risk analysis, when those conditions are protected.

5.3.1 The LACE Index. As a proof of concept, we evaluate the LACE index, which is a widely used method to predict 30-day readmission risk from EHR data [48]. LACE is a data driven method, which takes into account several clinical factors, including length of stay (L); acuity of the admission (A); comorbidity of the patient (C); and emergency visits in the six months prior admission (E), to compute risk score for each admission. Similarly to previous studies [25, 46, 52], we aim at classifying admission in three risk groups: low (score=0-4), moderate (score=5-9), and high (score $\geq$10) risk. Specifically, we compare the classification results obtained on the sanitization data with the ground truth labels obtained in the original data. The classification results are reported in Table 1.

The suppression method achieves privacy by redacting sensitive elements from the shared data, which may result in lower LACE index scores, as sensitive comorbidity may be removed. In our evaluations, we observe that the suppression method may misclassify high risk patients into a lower risk group, as we can see from false positive and false negative. As an example, 34 admissions at high risk are being classified as moderate risk. Together with the classification results, we also report the maximum attacker's precision (MAP). We observe that suppression does not provide adequate privacy protection, as the attacker's maximum precision is 1.0.

Our sanitization method achieves privacy via randomization, where the privacy parameter ϵ controls the level of privacy protection. With our approach, we observe that the classification results improve (i.e., higher true positive/negative and lower false positive/negative) as the values of privacy parameter increase (weaker privacy). With $\epsilon \geq 2.5$, we observe that for moderate and high risk groups, our approach achieves higher true positive and lower false negative compared to suppression, with limited false positive. As an example, for the high risk admissions our approach achieves: 33 TP, 11 FN, 5 FP, and 11697 TN, compared to 10 TP, 34 FN, 0 FP, and 11702 TN achieved by suppression. Moreover, we observe that our solution provides stronger privacy protection compared to suppression, with the attacker's maximum precision at most 0.63 for the weakest privacy setting.

Table 1: LACE index classification results with three class labels: low, moderate, and high risk, compared to the original data. MAP denotes the maximum precision for the inference attack observed on the sanitized data.

Privacy Method	Label	TP	FN	FP	TN
Suppression MAP = 1	Low	8844	0	292	2610
	Mod	2566	292	34	8854
	High	10	34	0	11702
ClinSimRR (ϵ = 1.0) MAP = 0.428	Low	8619	225	389	2513
	Mod	2463	395	262	8626
	High	7	37	6	11696
ClinSimRR (ϵ = 2.5) MAP = 0.429	Low	8672	172	333	2569
	Mod	2522	336	205	8683
	High	11	33	3	11699
ClinSimRR (ϵ = 5.0) MAP = 0.625	Low	8769	75	194	2708
	Mod	2659	199	86	8802
	High	33	11	5	11697

Table 2: Accuracy and recall for the readmission prediction task using neural networks. We apply the obfuscation approaches (Suppression and ClinSimRR) on the training set. For the non private approach, we trained the model on the original data with a percentage of patients with sensitive conditions opting-out from the training set. A balanced test set was used to generate the results.

Method	Accuracy	Recall
Suppression	0.77 (0.77,0.78)	0.41 (0.39,0.42)
ClinSimRR (ϵ = 2.5)	0.78 (0.77,0.78)	0.40 (0.38,0.42)
ClinSimRR (ϵ = 5.0)	0.77 (0.76,0.78)	0.41 (0.39,0.42)
NonPriv (opt%=0%)	0.77 (0.77,0.78)	0.42 (0.41,0.42)
NonPriv (opt%=50%)	0.78 (0.77,0.79)	0.40 (0.39,0.41)
NonPriv (opt%=100%)	0.79 (0.79,0.79)	0.37 (0.36,0.38)

5.3.2 Machine Learning Prediction. In addition to the LACE index, recent studies have proposed neural network models to predict readmission risks. Research networks and data consortium may leverage their large datasets to train highly accurate risk prediction models and then share these models with external sites. However, recent studies [42] have shown that models may leak information about data samples in the training set, posing significant privacy concerns. In the absence of adequate privacy protections, patients may be reluctant to participate in machine learning applications, potentially reducing the usability of the shared model.

We adapt the neural network model proposed recently by Lin et al. [36] for readmission risk prediction. Here we train the model using only diagnosis and procedure information, as our main goal is to assess the impact of the privacy mechanism on the predictive results provided by the model. Furthermore, we study the impact of individuals opting-out from the training set, when no privacy mechanisms are provided. Recall that patient authorization may be required for data sharing, as illustrated by Title 38 U.S.C. § 7332. To this end, we remove a varying percentage of patients with sensitive conditions from the training set (e.g., simulating patient opting-out). The results in terms of accuracy and recall are reported in Table 2. High values of recall indicate that the model is effective in determining patients at risk of readmission within 30 days.

We use 80% of the data (37064 patients) for training and construct a balance test set to assess the predictive results of the model. Among the patients in the training set, roughly 2300 individuals have sensitive conditions to be protected. While the fraction of patients with sensitive conditions is small compared to the overall data (6%), we observe that their contribution in the training set can benefit the performance of the model. As an example, we observe that the recall reduces from 0.42 to 0.37 as the fraction of patients with sensitive conditions opt-out from the training set, for the non-private solution. We argue that privacy mechanisms can be used to protect the entire training set, facilitating the retention of patients (especially those with sensitive conditions) in the training data. In this evaluation, the model trained on the privacy-enhanced data achieves recall and accuracy similar to those obtained with non-private and 0% opt-out.

6 RELATED WORKS

The problem of sharing protected health data has received much attention from the research community [19, 23]. A common privacy practice relies on de-identification, in which original data is manipulated to hide the identity of individual data contributors (e.g., suppression of PHI). Classic techniques for data de-identification

are build on the notion of k-anonymity [44], which have been conducted for clinical (e.g., EHR) and genomic data [17, 18, 38]. Recent approaches rely on data synthesis methods to generate fake records that highly resemble the original data [4, 6, 12, 50, 53]. Among them, Choi et al. [12] have proposed a data synthesis approach based on generative adversarial networks for EHR data. While these approaches may overcome some privacy issues by sharing synthetically generated data, they do not provide provable privacy protection, potentially leading to privacy vulnerability in the presence of an informed adversary. To provide rigorous privacy guarantees, recent studies have adopted the differential privacy model [16]. As an example, recent solutions leverage advanced machine learning techniques (e.g., deep learning models) and differential privacy for combined benefit, learning useful patterns from the original data in a privacy-protecting manner which are used to generate high quality individual-level data for secondary analysis [7, 34]. Despite promising results, these solutions provide protection at record-level (i.e., protecting the presence/absence of individual data contributors), thus they cannot support fine-grained privacy control (i.e., protecting specific elements in the record). In addition to data manipulation and synthesis, approaches based on security techniques have been proposed to support access-control and regulating information access [3, 26, 39, 43]. While these solutions provide an enhanced privacy control on data access, they are difficult to deploy in supporting broad data sharing.

In this work, we are interested in enabling fine-grained privacy control to support broad data sharing. Our setting differs from traditional privacy solutions in health applications, as we focus on protecting sensitive elements that represent a small subset of the overall data domain. Techniques based on suppression, generalization, and permutation have been proposed to hide sensitive data in data sharing applications [1, 8, 22, 51]. One challenge for these solutions is the data correlation, which may lead to privacy breaches. In our proposed approach, we aim at addressing the data correlation challenge by protecting dependent elements. Additionally, we leverage health domain knowledge to design an obfuscation mechanism that hides sensitive elements while retaining the clinical usefulness of the original data.

7 DISCUSSION

Sharing high quality sanitized health data is central in facilitating secondary analysis and supporting reproducible research. Yet, privacy solutions may diminish the usability of the shared data. Compared to traditional privacy approaches (e.g., standard differential privacy), our solution takes into account the domain-specific knowledge (i.e., ICD hierarchical representation) in the mechanism design to better retain data utility at individual level (i.e., codes). While our approach shows promising results, finding the right balance between privacy and utility is still an open problem. Some studies have proposed extensions of the differential privacy notion to defend against realistic adversarial models and to improve data usefulness in spatial and time-series analysis [5, 10, 20]. Studying the applicability of these relaxed privacy models in health applications can provide important insights to better understand realistic privacy risks.

The effectiveness of our approach in mitigating privacy breaches from dependent elements relies on how well data dependence can be modeled. In this work, we use a Bayesian network to model dependent elements in health data. Alternative approaches could be considered to better suit different application domains (e.g., mobile or location data). Overall, it may be helpful to consider two possible cases when data dependence is estimated from historic or public datasets. First, if the learned dependence overestimates what the adversary may know, our approach suffices to provide privacy protection. Second, if the learned dependence is weaker than what is known by the adversary, the adversary may leverage the dependent elements in the shared data to infer sensitive information, potentially breaching privacy. Nonetheless, our approach would provide an additional layer of privacy protection compared to traditional solutions (i.e., suppression, standard differential privacy). Studying the implications of dependent data on privacy protection is an important research topic that requires further investigations [11, 37, 54].

Enabling fine-grained privacy protection can help address specific privacy concerns of data contributors and ultimately facilitate research data participation [29, 32]. In this study, we investigate the possibility of enabling fine-grained privacy protection via algorithmic methods. However, solving such a problem goes beyond the development of privacy technology and the solution should be inclusive of human perception of privacy, legal factors, and ethical considerations. Overall, fulfilling individual privacy preferences (e.g., protecting sensitive elements), and providing rigorous, fine-grained privacy control has the potential for encouraging data participation and data sharing.

8 CONCLUSION

In this work, we studied the problem of sharing health data while providing fine-grained privacy protection. Our proposed approach leverages data dependence and domain knowledge to provide rigorous privacy protection while improving data utility. Empirical evaluation on real-world health data demonstrates the benefits of our approach compared to current privacy practices such as suppression and UMAXRR. Guided by the Discussions, future work may pursue several important directions: 1) future research could investigate the trade-off between privacy and utility with practical adversarial models (as opposed to the powerful adversary assumed by standard differential privacy); 2) future research could advance the modeling of data dependence in healthcare domain, potentially considering multi-modal data; 3) it is important to understand how fine-grained privacy control interacts with human perception of privacy, legal requirements, and ethical considerations.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers for their valuable feedback. LB is supported in part by National Human Genome Research Institute grants R00HG010493 and RM1HG009034, and a National Library of Medicine grant R01LM013712. LF is supported in part by National Science Foundation grants CNS-1951430 and CNS-2144684. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Osman Abul, Francesco Bonchi, and Fosca Giannotti. 2010. Hiding sequential and spatiotemporal patterns. *IEEE Transactions on Knowledge and Data Engineering* 22, 12 (2010), 1709–1723.

[2] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.

[3] Rakesh Agrawal and Christopher Johnson. 2007. Securing electronic health records without impeding the flow of information. *International journal of medical informatics* 76, 5-6 (2007), 471–479.

[4] Tanbir Ahmed, Md Momin Al Aziz, and Noman Mohammed. 2020. De-identification of electronic health record using neural network. *Scientific reports* 10, 1 (2020), 1–11.

[5] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 901–914.

[6] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. 2019. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association* 26, 3 (2019), 228–241.

[7] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. 2019. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes* 12, 7 (2019), e005122.

[8] Luca Bonomi, Liyue Fan, and Hongxia Jin. 2016. An information-theoretic approach to individual sequential data sanitization. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 337–346.

[9] Omar Bouhaddou, Jamie Bennett, Tim Cromwell, Graham Nixon, Jennifer Teal, Mike Davis, Robert Smith, Linda Fischetti, David Parker, Zachary Gillen, et al. 2011. The department of Veterans Affairs, department of defense, and Kaiser permanente nationwide health information network exchange in San diego: Patient selection, consent, and identity matching. In *AMIA Annual Symposium Proceedings*, Vol. 2011. American Medical Informatics Association, 135.

[10] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 82–102.

[11] Rui Chen, Benjamin CM Fung, S Yu Philip, and Bipin C Desai. 2014. Correlated network data publication via differential privacy. *The VLDB Journal* 23, 4 (2014), 653–676.

[12] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*. PMLR, 286–305.

[13] Roxana Dogaru, Flavia Micota, and Daniela Zaharie. 2015. Taxonomy-based dissimilarity measures for profile identification in medical data. In *2015 IEEE 13th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, 149–154.

[14] Margaret Donahue, Omar Bouhaddou, Nelson Hsing, Todd Turner, Glen Crandall, Joseph Nelson, and Jonathan Nebeker. 2018. Veterans health information exchange: successes and challenges of nationwide interoperability. In *AMIA Annual Symposium Proceedings*, Vol. 2018. American Medical Informatics Association, 385.

[15] John C Duchi, Michael I Jordan, and Martin J Wainwright. 2013. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 429–438.

[16] Cynthia Dwork. 2006. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*. Springer, 1–12.

[17] Khaled El Emam and Fida Kamal Dankar. 2008. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association* 15, 5 (2008), 627–637.

[18] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, et al. 2009. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association* 16, 5 (2009), 670–682.

[19] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. 2011. A systematic review of re-identification attacks on health data. *PloS one* 6, 12 (2011), e28071.

[20] Liyue Fan and Luca Bonomi. 2018. Time series sanitization with metric-based privacy. In *2018 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 264–267.

[21] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*. 17–32.

[22] Aris Gkoulalas-Divanis and Grigorios Loukides. 2011. Revisiting sequential pattern hiding to enhance utility. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1316–1324.

[23] Aris Gkoulalas-Divanis, Grigorios Loukides, and Jimeng Sun. 2014. Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of biomedical informatics* 50 (2014), 4–19.

[24] Michaela Götz, Suman Nath, and Johannes Gehrke. 2012. Maskit: Privately releasing user context streams for personalized mobile applications. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 289–300.

[25] Andrea Gruneir, Irfan A Dhalla, Carl van Walraven, Hadas D Fischer, Ximena Camacho, Paula A Rochon, and Geoffrey M Anderson. 2011. Unplanned readmissions after hospital discharge among patients identified as being at high risk for readmission using a validated predictive algorithm. *Open Medicine* 5, 2 (2011), e104.

[26] Cheng Guo, Ruhan Zhuang, Yingmo Jie, Yizhi Ren, Ting Wu, and Kim-Kwang Raymond Choo. 2016. Fine-grained database field search using attribute-based encryption for e-healthcare clouds. *Journal of medical systems* 40 (2016), 1–8.

[27] Zheng Jia, Xudong Lu, Huilong Duan, and Haomin Li. 2019. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC medical informatics and decision making* 19, 1 (2019), 1–11.

[28] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[29] Christine LM Joseph, Amy Tang, David W Chesla, Mara M Epstein, Pamala A Pawloski, Alan B Stevens, Stephen C Waring, Brian K Ahmedani, Christine C Johnson, and Cathryn D Peltz-Rauchman. 2022. Demographic differences in willingness to share electronic health records in the All of Us Research Program. *Journal of the American Medical Informatics Association* 29, 7 (2022), 1271–1278.

[30] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. 2016. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*. PMLR, 2436–2444.

[31] Daniel Kifer and Ashwin Machanavajjhala. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. 193–204.

[32] Jihoon Kim, Hyeoneui Kim, Elizabeth Bell, Tyler Bath, Paulina Paul, Anh Pham, Xiaoqian Jiang, Kai Zheng, and Lucila Ohno-Machado. 2019. Patient perspectives about decisions to share medical data and biospecimens for research. *JAMA network open* 2, 8 (2019), e199550–e199550.

[33] Ios Kotsogiannis, Stelios Doudalis, Sam Haney, Ashwin Machanavajjhala, and Sharad Mehrotra. 2020. One-sided Differential Privacy. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. 493–504. https://doi.org/10.1109/ICDE48307.2020.00049

[34] Dongha Lee, Hwanjo Yu, Xiaoqian Jiang, Deevakar Rogith, Meghana Gudala, Mubeen Tejani, Qiuchen Zhang, and Li Xiong. 2020. Generating sequential electronic health records using dual adversarial autoencoder. *Journal of the American Medical Informatics Association* 27, 9 (2020), 1411–1419.

[35] Sandra S-J Lee, Mildred K Cho, Stephanie A Kraft, Nina Varsava, Katie Gillespie, Kelly E Ormond, Benjamin S Wilfond, and David Magnus. 2019. "I don't want to be Henrietta Lacks": diverse patient perspectives on donating biospecimens for precision medicine research. *Genetics in medicine* 21, 1 (2019), 107–113.

[36] Chaohsin Lin, Shuofen Hsu, Hsiao-Feng Lu, Li-Fei Pan, and Yu-Hua Yan. 2021. Comparison of back-propagation neural network, LACE index and hospital score in predicting all-cause risk of 30-Day readmission. *Risk Management and Healthcare Policy* (2021), 3853–3864.

[37] Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. 2016. Dependence Makes You Vulnberable: Differential Privacy Under Dependent Tuples.. In *NDSS*, Vol. 16. 21–24.

[38] Grigorios Loukides, Aris Gkoulalas-Divanis, and Bradley Malin. 2010. Anonymization of electronic medical records for validating genome-wide association studies. *Proceedings of the National Academy of Sciences* 107, 17 (2010), 7898–7903.

[39] Kenneth D Mandl, William W Simons, William CR Crawford, and Jonathan M Abbett. 2007. Indivo: a personally controlled health record for health information exchange and communication. *BMC medical informatics and decision making* 7, 1 (2007), 1–10.

[40] Takao Murakami and Yusuke Kawamoto. 2019. Utility-optimized local differential privacy mechanisms for distribution estimation. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 1877–1894.

[41] Sunil Nair, Douglas Hsu, and Leo Anthony Celi. 2016. Challenges and opportunities in secondary analyses of electronic health record data. *Secondary Analysis of Electronic Health Records* (2016), 17–26.

[42] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.

[43] William W Simons, Kenneth D Mandl, and Isaac S Kohane. 2005. The PING personally controlled electronic medical record system: technical architecture. *Journal of the American Medical Informatics Association* 12, 1 (2005), 47–54.

[44] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems* 10, 05 (2002), 557–570.

[45] Stefano Taddei and Bastianina Contena. 2013. Privacy, trust and control: Which relationships with online self-disclosure? *Computers in human behavior* 29, 3 (2013), 821–826.

[46] Shu Yun Tan, Lian Leng Low, Yong Yang, and Kheng Hock Lee. 2013. Applicability of a previously validated readmission predictive index in medical patients in Singapore: a retrospective study. *BMC Health Services Research* 13, 1 (2013), 1–6.

[47] Hsiao-Ting Tseng, Fahad Ibrahim, Nick Hajli, Tahir M Nisar, and Haseeb Shabbir. 2022. Effect of privacy concerns and engagement on social support behaviour in online health community platforms. *Technological Forecasting and Social Change* 178 (2022), 121592.

[48] Carl Van Walraven, Irfan A Dhalla, Chaim Bell, Edward Etchells, Ian G Stiell, Kelly Zarnke, Peter C Austin, and Alan J Forster. 2010. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Cmaj* 182, 6 (2010), 551–557.

[49] Daniel M Walker, Sharona Hoffman, and Julia Adler-Milstein. 2022. Interoperability in a post-Roe era: sustaining progress while protecting reproductive health information. *JAMA* 328, 17 (2022), 1703–1704.

[50] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association* 25, 3 (2018), 230–238.

[51] Di Wang, Yeye He, Elke Rundensteiner, and Jeffrey F Naughton. 2013. Utility-maximizing event stream suppression. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. 589–600.

[52] Hao Wang, Richard D Robinson, Carlos Johnson, Nestor R Zenarosa, Rani D Jayswal, Joshua Keithley, and Kathleen A Delaney. 2014. Using the LACE index to predict hospital readmissions in congestive heart failure patients. *BMC cardiovascular disorders* 14, 1 (2014), 1–8.

[53] Chao Yan, Ziqi Zhang, Steve Nyemba, and Bradley A Malin. 2020. Generating electronic health records with multiple data types and constraints. In *AMIA annual symposium proceedings*, Vol. 2020. American Medical Informatics Association, 1335.

[54] Emre Yilmaz, Tianxi Ji, Erman Ayday, and Pan Li. 2021. Genomic Data Sharing under Dependent Local Differential Privacy. *arXiv preprint arXiv:2102.07357* (2021).

[55] Keyu Zhu, Pascal Van Hentenryck, and Ferdinando Fioretto. 2021. Bias and variance of post-processing in differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11177–11184.