Journal of the American Medical Informatics Association, 29(7), 2022, 1152–1160
https://doi.org/10.1093/jamia/ocac047
Advance Access Publication Date: 5 April 2022
Research and Applications

Research and Applications

Sharing personal ECG time-series data privately

Luca Bonomi[1], Zeyun Wu[2], and Liyue Fan[3]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [2]Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, California, USA, and [3]Department of Computer Science, University of North Carolina at Charlotte, Charlotte, North Carolina, USA

Corresponding author: Luca Bonomi, PhD, Department of Biomedical Informatics, Vanderbilt University, 2525 West End Avenue, Suite 1475, Nashville, TN 37203, USA; luca.bonomi@vumc.org

Received 9 September 2021; Revised 16 March 2022; Editorial Decision 17 March 2022; Accepted 31 March 2022

Objective: Emerging technologies (eg, wearable devices) have made it possible to collect data directly from individuals (eg, time-series), providing new insights on the health and well-being of individual patients. Broadening the access to these data would facilitate the integration with existing data sources (eg, clinical and genomic data) and advance medical research. Compared to traditional health data, these data are collected directly from individuals, are highly unique and provide fine-grained information, posing new privacy challenges. In this work, we study the applicability of a novel privacy model to enable individual-level time-series data sharing while maintaining the usability for data analytics.

Methods and materials: We propose a privacy-protecting method for sharing individual-level electrocardiography (ECG) time-series data, which leverages dimensional reduction technique and random sampling to achieve provable privacy protection. We show that our solution provides strong privacy protection against an informed adversarial model while enabling useful aggregate-level analysis.

Results: We conduct our evaluations on 2 real-world ECG datasets. Our empirical results show that the privacy risk is significantly reduced after sanitization while the data usability is retained for a variety of clinical tasks (eg, predictive modeling and clustering).

Discussion: Our study investigates the privacy risk in sharing individual-level ECG time-series data. We demonstrate that individual-level data can be highly unique, requiring new privacy solutions to protect data contributors.

Conclusion: The results suggest our proposed privacy-protection method provides strong privacy protections while preserving the usefulness of the data.

Key words: data privacy, ECG data, time-series, data sharing, predictive analytics

INTRODUCTION

Advances in mobile sensor technology (eg, wearable devices) are revolutionizing the way in which patient data are collected, enabling high-quality and fine-grained data to be gathered directly from personal devices.[1] These aggregated data have provided great opportunities for performing new analytics and advancing healthcare. For example, the use of wearable sensor data has been shown to be effective in facilitating the diagnosis, prevention, management of chronic diseases, and improving patient care.[2–4] Additionally, recent studies have shown that data collected from wearable devices could be used in the early detection of COVID-19.[5,6] Therefore, sensor data are increasingly integrated into large datasets (eg, NIH All of Us),[7] with the promise of providing the medical research community with new insights. However, sharing these sensor data broadly poses novel privacy challenges. First, the sole removal of personal identifiable information (eg, SSN) does not provide adequate privacy protection. Research studies have shown that individual-level sensor data could be used as biometric to identify individuals.[8,9] Specifically, Biel et al[9] have demonstrated that carefully selected features from electrocardiography (ECG) data could be used to achieve

100% reidentification rate on a dataset with 20 patients. Second, current privacy-protecting data sharing solutions require a trusted site to collect and manage the data (eg, central privacy model). However, individuals may lack trust in the data aggregator that has access to their original data.[10,11] To promote data participation and sharing, it is imperative to develop privacy methods that address the privacy needs of individual data contributors.

Current solutions for individual-level sensor data (eg, time-series) build on security and data anonymization primitives. Security-based solutions rely on encryption and access control techniques.[12–14] Despite promising results in some settings,[15–17] their paradigm allows only a small number of authorized users to access the data. Additionally, encryption solutions may still be vulnerable in the presence of an informed adversary (eg, in genomic applications[18]). To broaden data access, solutions based on *k*-anonymity[19] and differential privacy[20] have been proposed, which enable a data curator to sanitize the collected data and share the results with external users (eg, researchers).[21–24] However, these privacy solutions build on a traditional central authority assumption, in which a trusted data curator is responsible for collecting, aggregating, and protecting the individual health data. As a result, these traditional data anonymization approaches have limited applicability when individuals may not trust the data curator.

The objective of this work is to assess the privacy risks in sharing ECG time-series data and to develop novel individual-level privacy solutions to enable individuals to share data with an untrusted data aggregator. By providing privacy at individual-level, data contributors have greater control over their private data compared to the traditional central privacy models, which may facilitate data participation. We show that the ECG time-series data considered in this study are highly unique, enabling an informed adversary to perform accurate inference attacks. One example attack is membership inference: an individual who participates to a research study may inadvertently disclose partial ECG time-series data (eg, she may share her ECG measurements during a certain day with other mobile applications). An adversary may leverage the disclosed information together with the aggregated data shared by the study to infer the presence of the target individual. To mitigate these privacy risks, we propose a privacy-protecting method that achieves strong privacy protection, enabling individuals to sanitize their data as they are collected for research studies. Our method builds on the metric privacy model,[25,26] a generalization of the rigorous differential privacy notion, which provides provable privacy protection for individual-level data. Our privacy protecting solution can be deployed directly on the sensor devices (eg, wearables and smartphones), protecting the ECG time-series data as they are collected. The evaluation shows that our solution is effective in providing privacy protection against an informed adversary, significantly reducing known privacy risks (eg, membership and attribute inference). We also show that the sanitized aggregate-level data can enable accurate health analytics (eg, predictive and clustering tasks), demonstrating that the sanitized data preserve the clinical usefulness of the original data.

MATERIALS AND METHODS

Application setting

In a mobile sensing application, ECG data are generated by n individuals and are sent to an untrusted data aggregator, which aggregates these data and shares them with researchers (Figure 1). On the client side, each individual applies our privacy method to the original ECG time-series data S to generate a sanitized time-series $\widehat{S}$. These sanitized data are shared with the data aggregator, where a pair $(ID,\ \widehat{S})$, represents a pseudo ID and the sanitized ECG data contributed by an individual. On the data aggregator side, the ECG data may be associated with a label (eg, clinical data linked to the pseudo ID) and the aggregated ECG and label data are shared with external researchers to enable predictive tasks (eg, cardiovascular disease classification).

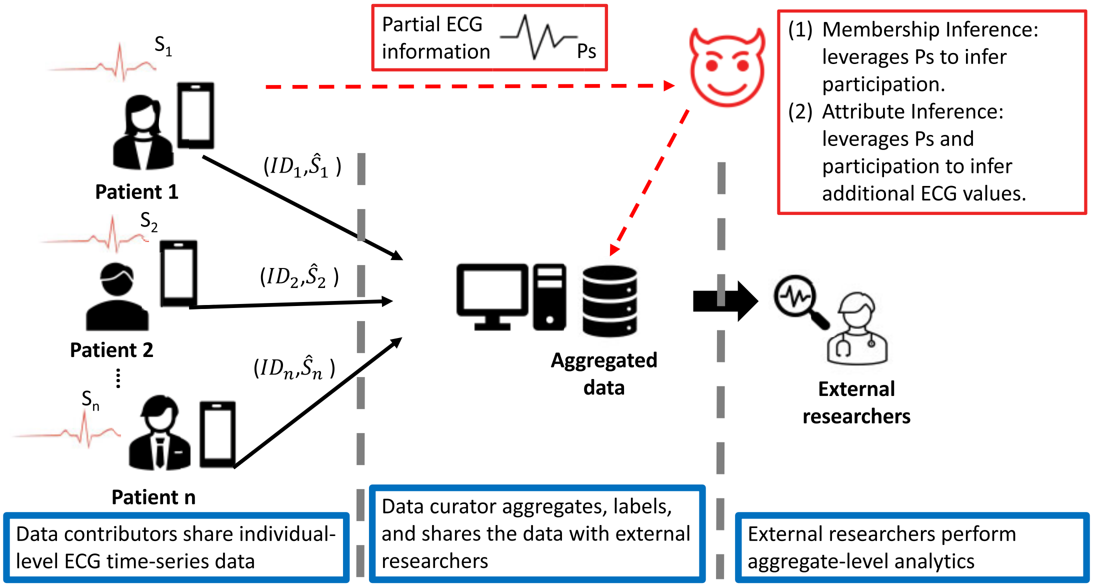
S1
Partial ECG information
Ps
(1) Membership Inference: leverages Ps to infer participation.
(2) Attribute Inference: leverages Ps and participation to infer additional ECG values.
Patient 1
(ID1,Ŝ1)
S2
(ID2,Ŝ2)
Patient 2
(IDn,Ŝn)
Sn
Patient n
Aggregated data
External researchers
Data contributors share individual-level ECG time-series data
Data curator aggregates, labels, and shares the data with external researchers
External researchers perform aggregate-level analytics

Figure 1. Overview of the application setting. Our solution is deployed on each device enabling individuals to protect the original ECG time-series data and to share only a sanitized version of their data. The sanitized data are collected and aggregated by a central data aggregator, which may share the aggregated data with external researchers. We consider 2 types of attacks that may lead to privacy breaches. (1) An adversary may have access to partial information about the original time-series data of a target individual and may leverage the sanitized data to determine whether the target contributed to the study. (2) An adversary who knows the participation of a target individual may leverage the sanitized data to infer sensitive attributes. Our proposed solution can be deployed on each individual wearable device, enabling the collection of sanitized time-series ECG data while protecting privacy.

The application setting is depicted in Figure 1 and presents 2 major privacy challenges. (1) An adversary who has access to known ECG information about a target may leverage the shared data to conduct membership inference attacks with the goal of determining whether the target contributed to the data. (2) An adversary who has access to known ECG information about a target as well as the target's participation in the data may infer additional sensitive time-series values (eg, attribute inference). The goal of our work is to quantify and mitigate these privacy risks. Furthermore, there is a tension between protecting individual privacy and the usability of the data: privacy solutions need to preserve usability of the aggregated time-series data to support research studies. Therefore, it is imperative to provide individuals with strong privacy control over the shared individual-level data while preserving the usability at aggregate-level.

Sanitizing individual time-series data

Here, we propose a data sanitization method to provide rigorous privacy for individual-level ECG time-series data. Our method allows individuals to share protected time-series ECG data $\widehat{S}$, while the original time-series S never leave the individual's device. Our sanitization method builds on *metric privacy*,[26] which is a generalization of the differential privacy model and provides provable privacy protection for individual-level data.[27,28]

Metric privacy model

The standard differential privacy model provides indistinguishability for any pair of databases D and D' that differ in any single record. In contrast, the metric privacy model generalizes the notion of indistinguishability by considering the difference between input values captured with a distance metric d_X. Specifically, given an arbitrary set of secrets X (eg, databases) with a metric d_X, the metric privacy model is defined as follows.

Metric privacy (definition 1 in reference 26)

A mechanism $M: \ X \ \rightarrow P(Z)$ satisfies d_X-privacy, if and only if $\forall x, \ x' \in \ X$ the following inequality holds: $M(x)(Z) \ \le e^{d_X(x,x')}$ $M(x')(Z) \ \forall Z \in \mathcal{F}_Z$, where Z is a set of query outcomes and $P(Z)$ is the set of probability measures over Z.

Chatzikokolakis et al[26] have shown (in Proposition 1) that the metric d_X can be obtained by scaling a standard metric with a factor ε, which is the privacy parameter/budget. The authors[26] have also shown that standard differential privacy can be achieved via metric privacy by using the Hamming distance between datasets, ie, $d_X = \varepsilon \times d_H$. In recent years, several approaches have been proposed based on metric privacy.[27,29] Andrés et al[27] adopted the geographical distance between locations to provide indistinguishability in spatial databases. Their intuition is that it may be acceptable to reveal the approximate information about the individual's location (eg, the individual is in city A, not city B) while protecting the precise location (eg, the individual's exact address in city A). In addition, several theoretical privacy frameworks have been proposed, such as Blowfish[30] and Pufferfish[31] privacy, which exhibit similarity to our privacy definition. In the Supplementary Appendix, we describe these models and discuss how they relate to this work. Overall, the metric privacy model aims at providing a more generalizable privacy notion compared to the standard differential privacy, which could improve the usefulness of the shared data and benefit biomedical applications. However, it is challenging to define metrics that are meaningful from both clinical and privacy perspectives. To this end, we study how to apply the metric privacy model to ECG time-series data.

Metric privacy for ECG time-series

Our goal is to achieve indistinguishability between individual-level ECG time-series. Specifically, coarser information should be preserved to enable useful analytics, such as classifying normal heartbeat versus myocardial infarction, while fine-grained data are made indistinguishable to protect the privacy of individual data contributors. Given any pair of time-series S_1 and S_2, our method aims at bounding the ability of an adversary who observes the sanitized time-series $\widehat{S}$, to determine whether the input was S_1 or S_2. To this end, we apply the metric privacy model in the discrete cosine transform (DCT) domain. The DCT domain can well capture the structure of time-series data, in which the similarity between time-series can be estimated. Furthermore, the DCT coefficients serve as a compact representation, which can benefit both efficiency and privacy. Specifically, we consider the Euclidean distance between the first l coefficients in the DCT domain (ie, $d_2(DCT_l(S_1), DCT_l(S_2))$) to capture the distance between 2 time-series. Using this metric, we perform a private sampling to sanitize the vector representation of the input time-series (see Supplementary Appendix). Our previous work shows that the sampling mechanism satisfies d_X -privacy, with $d_X \ = \varepsilon \times d_2(DCT_l(S_1), DCT_l(S_2))$ and ε is the privacy parameter (Theorem 1, in reference[29]). After sampling, we perform the inverse DCT transformation with the sampled coefficients, generating a sanitized ECG time-series $\widehat{S}$.

In sensing applications, an individual may generate a continuous stream of ECG time-series data. For example, a wearable device may record a time-series for each heartbeat, generating a collection of time-series segments associated with the individual (ie, forming a profile for the patient). Because the total number of segments in each patient's profile may be unknown a priory, we protect each individual time-series segment independently in our sanitization method. While a higher privacy cost may be accumulated by this approach, it offers consistent and strong privacy protection to each time-series. This approach is also used in real-world applications (eg, Apple's privacy safeguard[32,33]), where the privacy budget is refreshed after a fixed amount of time (eg, 1 day). In principle, to provide a bounded overall privacy guarantee, we can divide the overall privacy budget by the total number of time-series generated by the individual.[26,27]

Empirical privacy measures

The parameter ε in the metric privacy model can be tuned to control the provable privacy protection, where lower values indicate stronger privacy. However, it is important to understand how this theoretical privacy protection mitigates practical privacy risks. A variety of privacy risk measures, including membership and attribute inference risks, have been proposed for well-studied health data types (eg, genomics, EHR, and tabular data).[11,34–39] In this work, we adapt those privacy measures to suit ECG time-series data based on realistic adversarial models. Below, we briefly describe our privacy measures, a detailed description is reported in the Supplementary Appendix.

Data uniqueness

Data uniqueness and privacy protection are closely related. As an example, several privacy methods rely on data manipulation techniques (eg, generalization) to reduce data uniqueness and achieve privacy (eg, k-anonymity[19]). In our work, we measure the uniqueness of the recorded ECG time-series both in the original and sanitized data. Our measure of data uniqueness is inspired by the privacy measure for mobility data proposed by De Montjoye et al.[40] In our setting, the uniqueness score $u_k(D)$ of the shared dataset D represents the fraction of individuals that can be uniquely identified by a subset of k time-series readings. Higher values of uniqueness score indicate that the time-series are more unique.

Membership inference

We consider an informed adversary who has some prior knowledge about the original time-series of a target and aims at determining whether the target participated in the sanitized data. An individual may inadvertently disclose partial information collected by her mobile sensor devices, for example, sharing sensor data with other applications or on social media. An adversary may leverage the known information of a target to determine whether the target contributed to the study with a distance-based approach, as described in recent works.[38,41] In our attack model, the adversary uses the dynamic time warping distance[42] (DTW) to match the known information of the target to the sanitized data. Then, using a threshold value th, the adversary determines that the target was included in the data if the target can be matched to a sanitized profile within DTW distance th. We measure the success of such attack in terms of accuracy, where higher values of accuracy indicate higher success in learning the membership of the target in the data.

Attribute inference

In attribute inference, the informed adversary knows partial information about the target's time-series data and their participation to the study. In fact, an individual may self-disclose their participation to a research study, for example, on social media.[43,44] The adversary may leverage the sanitized data to infer the value of the remaining time-series of the target. As an example, by inspecting the sanitized data the adversary could reconstruct a time-series in the target profile representing a sensitive heartbeat (eg, myocardial infarction), thus learning a sensitive condition of the target. In our attack model, the adversary uses a K-NN framework to impute the unknown time-series values. Specifically, the adversary first identifies the top-K profiles in the sanitized data similar to the known time-series, and then impute the unknown time-series values (eg, average among the ECG readings). The difference between the imputed and the original values (ie, inference error) can be used to quantify the adversary's ability to infer the unknown ECG readings of the target, where lower error values indicate higher accuracy in inferring the original values.

RESULTS

Settings

We use 2 publicly available datasets: ECG200 and BIDMC. ECG200 dataset is sampled from the MIT-BIH Supraventricular Arrhythmia Database.[45] The dataset comprises 200 ECG recordings of a single patient, each representing the electrical activity recorded during 1 heartbeat with 96 values. The time-series are labeled into 2 classes: normal heartbeat and a myocardial infarction.[46] BIDMC comprises ECG recordings, each of 8-min duration sampled at 125 Hz, from 53 ICU patients at the Beth Israel Deaconess Medical Center.[47,48] In this dataset, we divide the ECG time-series into segments of 0.64 s and consider the first 25 segments in each individual's profile. We set the number of DCT coefficients to $l = 24$ and vary the privacy parameter ε in the range $[1,\ 20]$. These parameters were selected according to the guidelines suggested in our previous privacy study for time-series data.[29]

Privacy evaluations

Data uniqueness

Figure 2 reports the values of uniqueness for the ECG time-series in the BIDMC dataset. To compute the uniqueness, we map the original real-valued time-series data into a discrete domain. Once discretized, we quantify the uniqueness of the time-series by considering subsequence of consecutive readings that can uniquely identify a time-series in the data. Figure 2(A) shows that for a domain size of 25 symbols, subsequences of 5 readings can uniquely identify more than 89% of the individuals in the BIDMC dataset. We notice that the uniqueness increases as more readings are used. Furthermore, as we decrease the domain size (ie, time-series are mapped into fewer symbols), the uniqueness decreases. Overall, the time-series in the original BIDMC data are highly unique. Figure 2(B) reports similar results on the time-series sanitized with our privacy method. While our privacy method

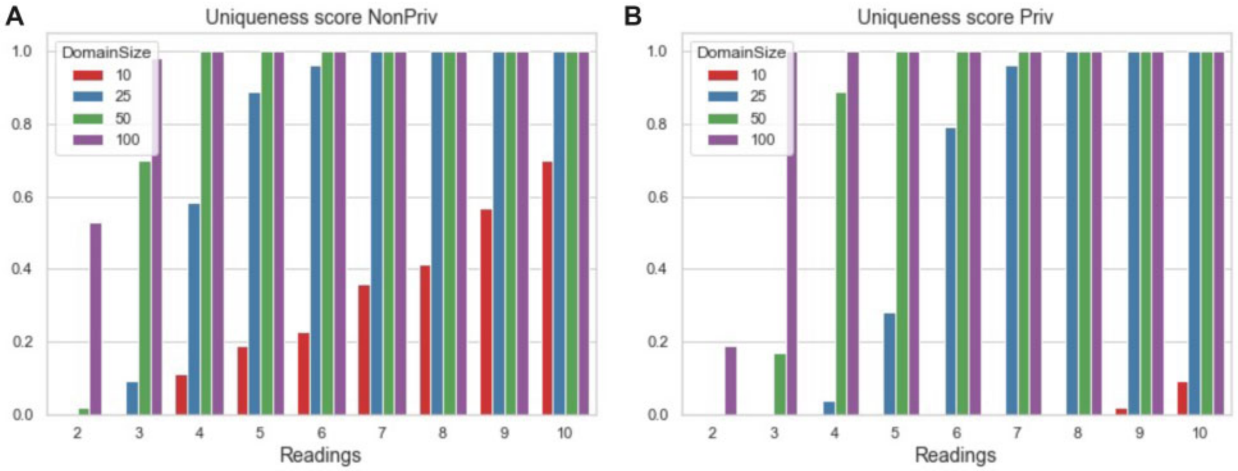

Figure 2. Empirical evaluation for the uniqueness of the ECG time-series data with different numbers of readings and domain size. Results are reported with 95% confidence intervals. (A) Uniqueness for the ECG time-series in the original BIDMC data (NonPriv). (B) Uniqueness for the ECG time-series in the sanitized BIDMC data (Priv) with $\varepsilon = 5.0$.

reduces the data uniqueness compared to the original data, we observe that the time-series are still unique for large domain sizes and higher numbers of readings. Nevertheless, we will show that the sanitized data are protected from membership and attribute inference attacks. In fact, high data uniqueness does not necessarily imply low privacy protection. Specifically, our randomized approach mitigates existing inference attacks by reducing the ability of the adversary to link the known data of the target with the output sanitized data.

Membership inference risk evaluations

Figure 3 reports the empirical membership inference risk with different parameter values. In the membership attack, the adversary relies on a threshold (*th*) to determine whether the known ECG profile of the target is in the sanitized data (see detailed methodology in Supplementary Appendix). To quantify the success of membership inference, we measure the accuracy for different threshold values. In original data (ie, NonPriv) the attack accuracy decreases as larger values of *th* are used. Figure 3(A) shows that the accuracy is high even when the adversary knows only 1 segment, and it increases with the attacker's background knowledge (ie, number of time-series segments in target's profile), where the highest value of accuracy is achieved when 25 time-series are considered. In Figure 3(B), we assess whether discretizing the time-series data into discrete, coarse representations may provide privacy protection. From these results, we observe that even with a coarser representation (eg, domain size 10) the adversary is successful in membership inference. In Figure 3(C), we observe that our sanitization method provides robust privacy protection against an adversary with increasing background knowledge about the target profile. In fact, the attack accuracy is always below 60%. In Figure 3(D), we report the impact of privacy parameter ε on the attack accuracy. Overall, the attack accuracy on the sanitized data is significantly reduced compared to nonprivate data. We report additional results on the attack precision and recall with different values of epsilon in the Supplementary Appendix, which show that our randomized method makes it challenging to associate the sanitized data with target profiles by tuning the threshold, thus leading to low attack accuracy.

Attribute inference risk evaluations

Figure 4 reports the average attribute inference error for the BIDMC dataset. In our attack model, the adversary uses the K-NN based

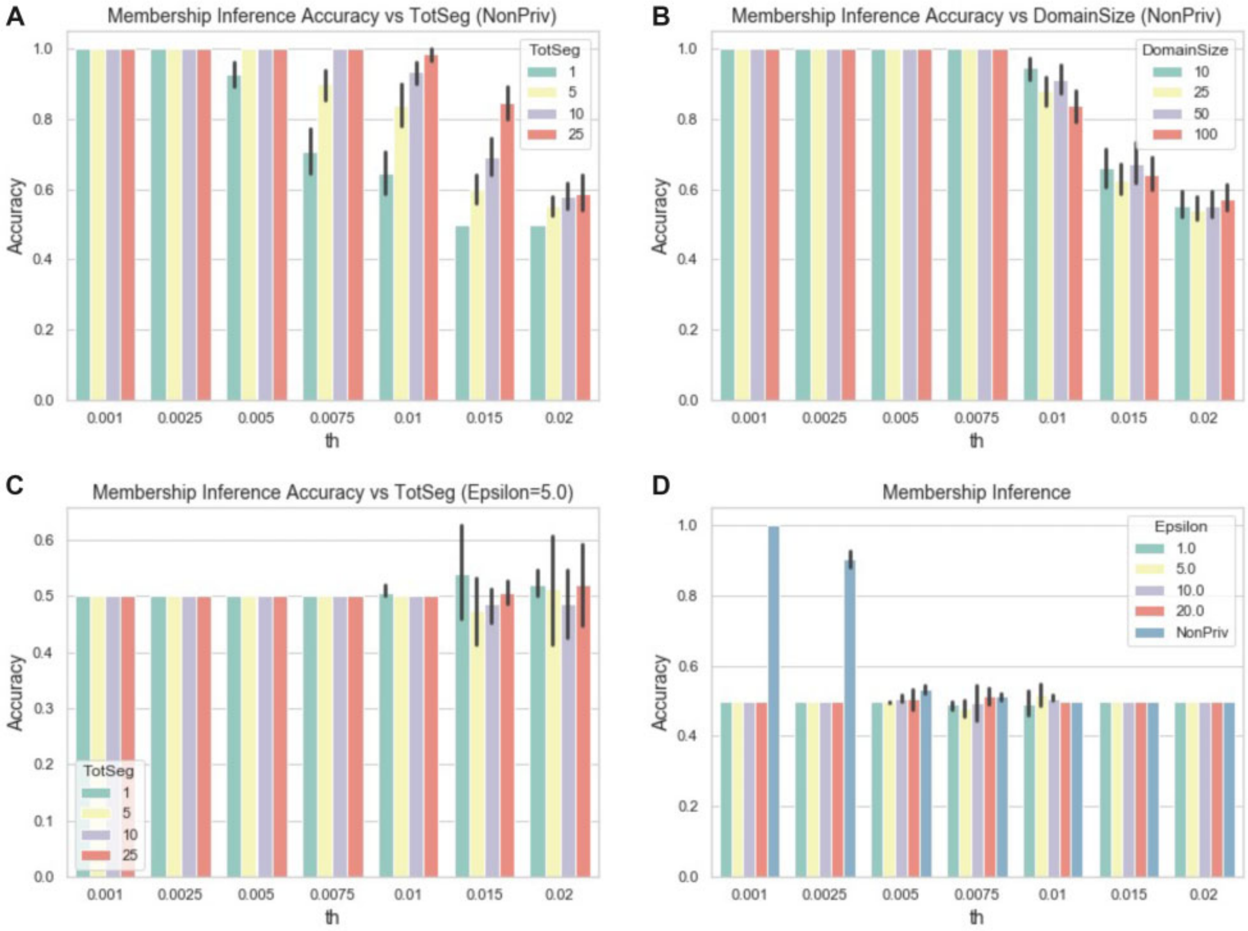

Figure 3. Empirical membership inference risk evaluation. The success of the adversary in inferring the membership of a target individual in the sanitized data is measured in terms of accuracy. We report the privacy measure both on the original data (NonPriv) and on the sanitized data with different values of the privacy parameter epsilon. Results are reported with 95% confidence intervals. (A) Accuracy of the membership inference attack on the original BIDMC (NonPriv) with an increasing number of time-series recorded for each individual (TotSeg). (B) Accuracy of the membership inference attack on the discretized BIDMC data (NonPriv) with different value of domain size. (C) Accuracy of the membership inference attack on the sanitized BIDMC data (Priv) with an increasing number of time-series recorded for each individual (TotSeg) and $\varepsilon = 5.0$. (D) Accuracy of the membership inference attack on the ECG200 dataset with different values of the privacy parameter.

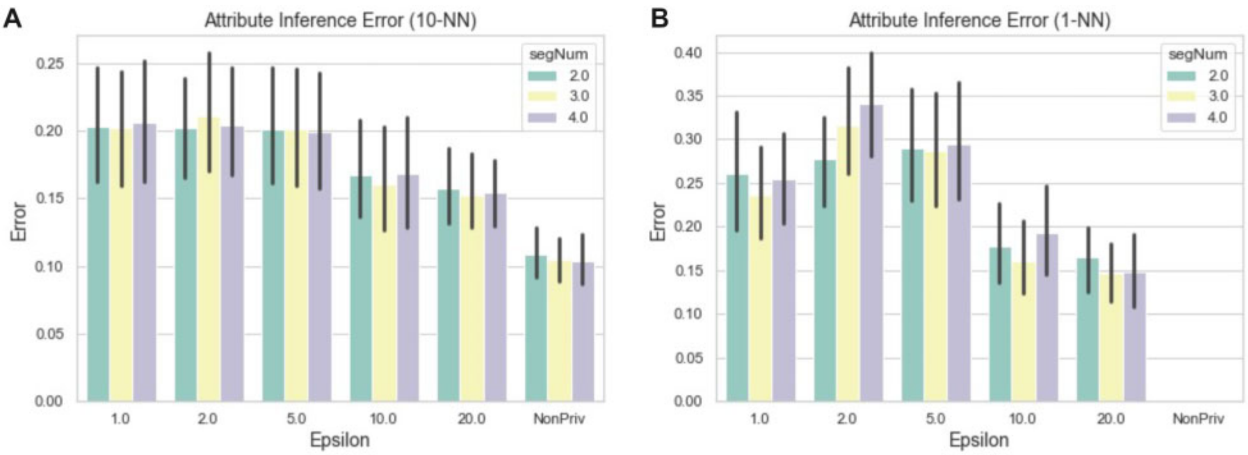

Figure 4. Empirical attribute inference error evaluation on the BIDMC dataset. The adversary aims at recovering unknown ECG readings shared by a target individual. In our evaluation, each individual generates a profile comprising 5 time-series. An informed adversary has access to some time-series of the target (segNum) and aims at recovering the values of the remaining time-series in the profile. Results are reported with 95% confidence intervals. (A) Attribute inference error using 10-NN. (B) Attribute inference error using 1-NN.

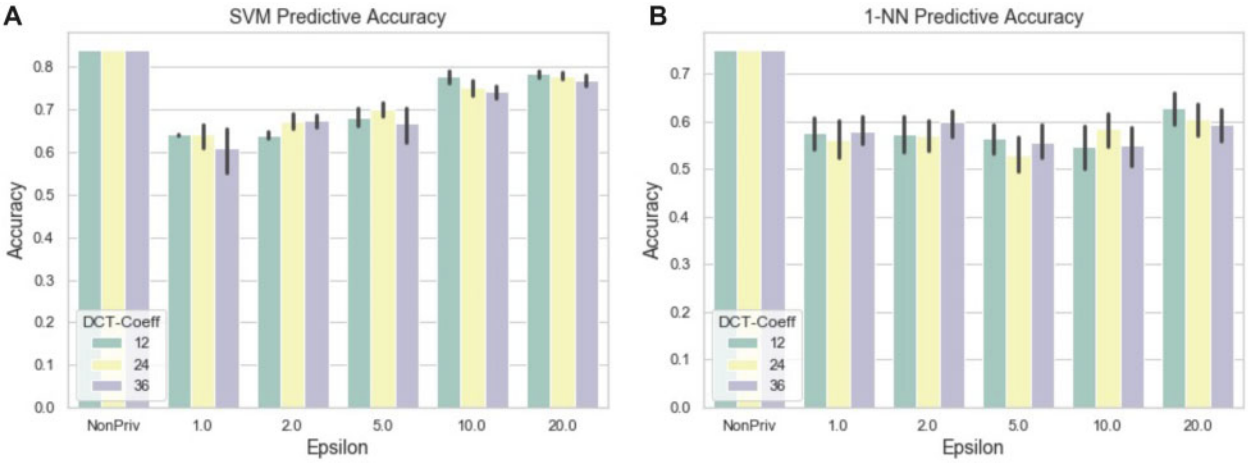

Figure 5. Utility evaluation in ECG classification. We report the accuracy for the classifiers on the nonprivate data and the accuracy results on the sanitized data with increasing values of the privacy parameter. Results are reported with 95% confidence intervals. (A) Accuracy results for the SVM classifier with increasing values of privacy parameter (ε) and different number of DCT coefficients. (B) Accuracy results for the 1-NN classifier with increasing values of privacy parameter (ε) and different number of DCT coefficients.

method to impute the unknown time-series shared by the target individual in the sanitized data. In this setting, we fix the number of time-series segments per profile to 5, and we assess the impact of the adversary's background knowledge on the attribute inference error by increasing the background knowledge of the adversary (ie, varying the number of time-series segments known by the adversary from 2 to 4). Additionally, we vary the number of neighboring profiles used in the imputation process. In the original data, the average inference error (ie, error in imputing the unknown readings) is roughly 10% for $K = 10$, while for $K = 1$ the adversary can correctly recover the unknown time-series of the target. Our proposed privacy method increases the inference error inflicted by the adversary for both $K = 10$ anid $K = 1$. As an example, with $\varepsilon = 5.0$, the average attribute inference error is roughly 20% and 28% for $K = 10$ and $K = 1$, respectively. The average error gently decreases as the privacy protection is relaxed. Additionally, we observe that in weaker privacy settings (ε =20 or nonprivate), an adversary with higher values of segNum achieves lower inference error. With stronger privacy, higher values of segNum do not necessarily lead to lower inference error, as the sanitized time-series may not accurately retain the DTW distance between profiles. Overall, our sanitization method significantly mitigates attribute inference attack on the sanitized data.

Usability of the shared ECG data

Classification results

To evaluate the usability of the sanitized data, we consider an ECG classification task using 2 classifiers: support vector machine (SVM) and the 1-nearest neighbors (1-NN). For the NN method, we use the DTW to measure the distance between time-series data. These classifiers are simple machine learning models that are widely used in healthcare application settings, including time-series classification.[49] In these evaluations, we use the ECG200 dataset, in which the recorded time-series data are labeled in 2 classes: normal heartbeat and myocardial infarction. The overall data are divided into 80% training and 20% testing. Our goal is to assess the impact of the privacy protection on the predictive accuracy of these classifiers. To evaluate the impact of privacy, we train the classifiers on the sanitized training set with increasing values of the privacy parameter. Then, we test the classifiers on the nonprivate test set. Figure 5 reports the predictive accuracy results on the test set with the classifiers trained on the original versus sanitized data. Among the 2 classifiers, SVM outperforms 1-NN with DTW. From Figure 5(A), we observe that the predictive accuracy for the SVM classifier approaches the results on the nonprivate data as epsilon increases. As an example, with $\varepsilon \geq 10.0$, the average predictive accuracy is above 75%. Regarding the results for the 1-NN classifier, the predictive accuracy results in Figure 5(B) are robust against changes in the privacy parameter, achieving predictive accuracy results above 60% for all settings.

Clustering results

On the BIDMC dataset, we use an unsupervised clustering approach to identify groups of individuals with similar ECG time-series profiles and assess whether the privacy method proposed in this work may affect the clustering results. Specifically, we consider a hierarchical clustering approach where the DTW distance is used to quantify the similarity between ECG profiles. In principle any unsupervised clustering approach could be used, and hierarchical clustering is a preferred method to eliminate randomness (eg, in comparison to k-means). Using the hierarchical clustering approach, we identify 2 distinct clusters among individuals in the original data. Information about the age and gender distribution in the identified clusters is summarized in Table 1. Then, we perform clustering on the sanitized data produced with different values of the privacy parameter. To visualize these high dimensional clusters, we use t-SNE technique,[50] in which the results are projected into a 2-dimensional space (Figure 6A and B). Additionally, we report all-pairs distance between ECG profiles in the original data and sanitized data (Figure 6C and D). As our privacy method performs random sampling in the DCT domain, it may incur utility loss for clustering sanitized profiles. We observe that for large epsilon values (eg, $\varepsilon = 20$), the pair-wise distance between ECG time-series is well preserved, enabling the hierarchical clustering method to generate clusters that well resemble those identified in the original data.

Table 1. Information about gender and age for the patients in hierarchical clustering with the DTW distance between their ECG time-series profiles

	Label 0 ($n = 12$)	Label 1 ($n = 41$)
Female	25%	56.1%
Male	75%	43.9%
Age < 65	50%	43.9%
Age ≥ 65	50%	56.1%

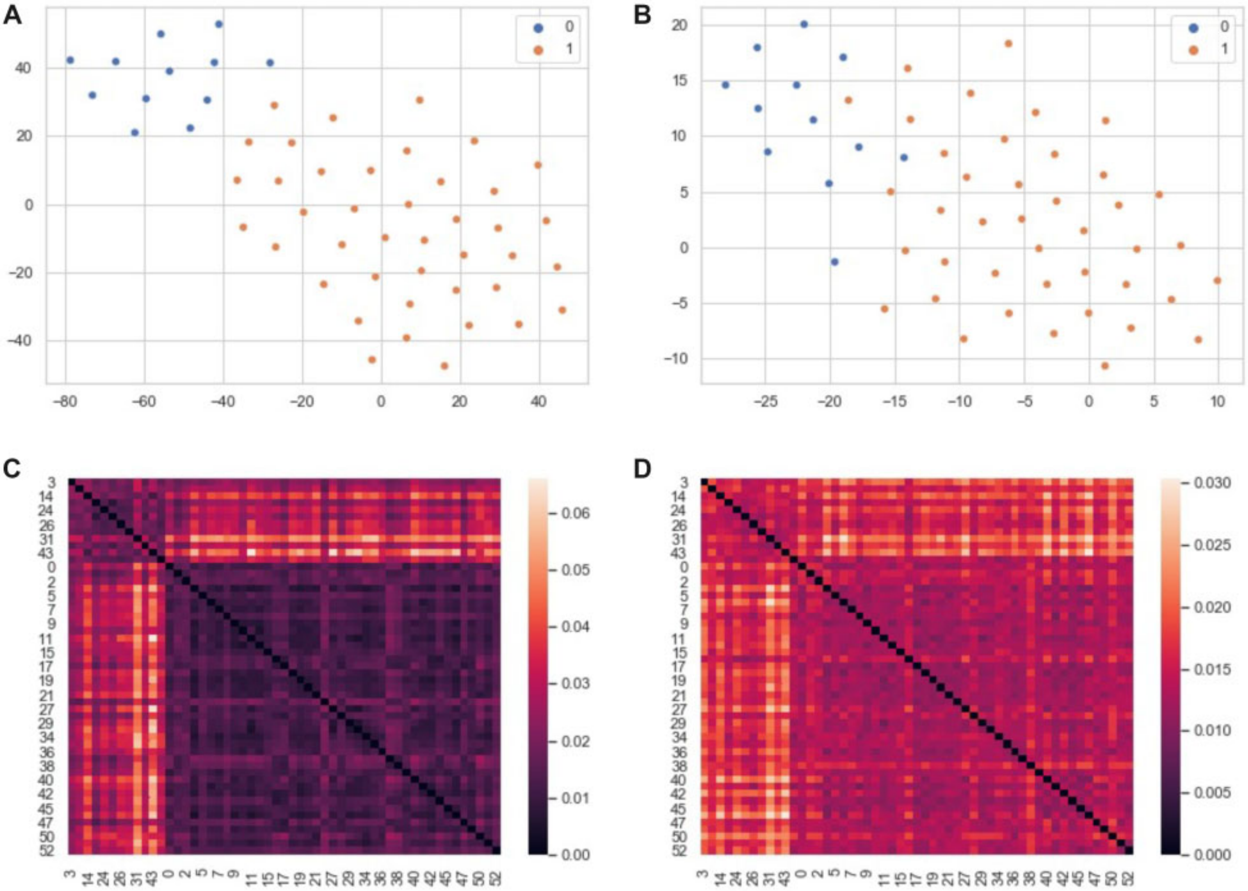

Figure 6. Distance-based clustering visualization on the BIDMC dataset. Using a hierarchical clustering approach, we identified 2 clusters (labeled 0 and 1) among all individuals. The clustering results for original data and sanitized data are visualized by t-SNE and are reported in (A) and (B), respectively. We observe that the general structure in clustering is preserved in the sanitized data. The all-pairs distance between ECG profiles on both the original and sanitized data are reported in (C) and (D), respectively. While the absolute distances between profiles are reduced in the sanitized data, the structure of distance matrix is preserved, which can be used to separate the clusters. The privacy parameter for obtaining the sanitized results was $\varepsilon = 20$.

DISCUSSION

Our evaluations provide important insights on the design of privacy-protecting methods that can be deployed to protect individual-level health data.

We observed that ECG time-series data are highly unique in the datasets considered in this study. Our results show that a coarser data granularity can help reducing the uniqueness in the data, but it may not prevent membership inference attacks. For example, deterministically reducing the domain size had limited effects on the attack accuracy of the informed adversary considered in this work (Figure 3B). Our proposed sanitization method uses random sampling to introduce uncertainty in the output time-series data, reducing the ability of the adversary to link the target's data with the sanitized output, thus reducing the success of membership inference.

In the differential privacy model, large values of the privacy parameter indicate weaker privacy protection. As a result, small privacy parameters (eg, $\varepsilon \leq 1.0$) may be used to provide strong privacy guarantees, at a high cost of data usability. In our evaluations, we show that our solution can successfully mitigate existing inference attacks even for large values of the privacy parameter, enabling useful analytics (eg, clustering and predictive tasks) otherwise impossible with small privacy parameters. Future research efforts in studying the gap between theoretical privacy guarantees and practical adversarial models could provide useful insights in the design of privacy mechanisms, in order to find the right balance between privacy and usability.

Recent advances in technology have made it possible to collect fine-grained data directly from individuals (eg, Direct-To-Consumer genomics, wearable devices). Although this manuscript addresses some of the privacy issues in individual-level data sharing, more research in the areas of ethics, human–computer interaction, and education is needed to advance the design of individual-level privacy solutions.

CONCLUSION

In this work, we studied the applicability of the formal metric privacy model to provide privacy protection for individual-level ECG time-series data. Our evaluations demonstrated that our sanitization approach can provide strong privacy protection against powerful privacy attacks, and the aggregate ECG data can be used to develop predictive models and fine-grained analysis for cardiovascular diseases. Overall, our privacy study provides important insights on the development of privacy-protecting pipelines for collecting individual-level data and making them available for secondary use, which could facilitate emerging health applications (eg, telemedicine and personalized medicine).

FUNDING

The National Human Genome Research Institute R00HG010493, National Library of Medicine grant R01LM013712, and by the National Science Foundation Award number 2040727 to LB, in part; the National Science Foundation CNS-1949217, CNS-1951430, and a UNC Charlotte FRG award to LF, in part. 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.

AUTHOR CONTRIBUTIONS

LB provided the motivation for this work, developed the method, contributed most of the writing, and conducted the experiments. ZW contributed with the usability evaluations and provided helpful comments. LF provided the motivation for this work, detailed edits, and critical suggestions.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

Authors declare no competing interests.

DATA AVAILABILITY

The datasets used in this article are freely accessible to interested researchers. More information can be found on the websites for BIDMC (https://physionet.org/content/chfdb/1.0.0/) and ECG200 (https://timeseriesclassification.com/description.php?Dataset=ECG200).

REFERENCES

1. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. *Per Med* 2018; 15 (5): 429–48.
2. Oresko JJ, Duschl H, Cheng AC. A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing. *IEEE Trans Inf Technol Biomed* 2010; 14 (3): 734–40.
3. Sim I. Mobile devices and health. *N Engl J Med* 2019; 381 (10): 956–68.
4. Uddin M, Syed-Abdul S. Data analytics and applications of the wearable sensors in healthcare: an overview. *Sensors* 2020; 20 (5): 1379.
5. Ates HC, Yetisen AK, Güder F, Dincer C. Wearable devices for the detection of COVID-19. *Nat Electron* 2021; 4 (1): 13–4.
6. Quer G, Radin JM, Gadaleta M, *et al.* Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nat Med* 2021; 27 (1): 73–7.
7. All of Us Research Program Investigators. The "All of Us" Research Program. *N Engl J Med* 2019; 381 (7): 668–76.
8. Irvine JM, Israel SA, Wiederhold MD, Wiederhold BK. A new biometric: human identification from circulatory function. In: joint statistical meetings of the American Statistical Association; 2003: 1957–63; San Francisco, CA, USA.
9. Biel L, Pettersson O, Philipson L, Wide P. ECG analysis: a new approach in human identification. *IEEE Trans Instrum Meas* 2001; 50 (3): 808–12.
10. Kim J, Kim H, Bell E, *et al.* Patient perspectives about decisions to share medical data and biospecimens for research. *JAMA Netw Open* 2019; 2 (8): e199550.
11. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. *Nat Genet* 2020; 52 (7): 646–9.
12. Sufi F, Khalil I. Enforcing secured ECG transmission for realtime telemonitoring: a joint encoding, compression, encryption mechanism. *Security Commun Netw* 2008; 1 (5): 389–405.
13. Sufi F, Khalil I, Jiankun H. ECG-based authentication In: Stavroulakis P, Stamp M, eds. *Handbook of Information and Communication Security*. Berlin: Springer; 2010: 309–31.
14. Layouni M, Verslype K, Sandıkkaya MT, De Decker B, Vangheluwe H. Privacy-preserving telemonitoring for ehealth. In: IFIP annual conference on data and applications security and privacy; July 12, 2009: 95–110; Montreal, Canada.
15. Poon CCY, Zhang Y-T, Bao S-D. A novel biometrics method to secure wireless body area sensor networks for telemedicine and m-health. *IEEE Commun Mag* 2006; 44 (4): 73–81.

16. Pandey S, Voorsluys W, Niu S, Khandoker A, Buyya R. An autonomic cloud environment for hosting ECG data analysis services. *Future Gener Comput Syst* 2012; 28 (1): 147–54.
17. Bhalerao S, Ansari IA, Kumar A, Jain DK. A reversible and multipurpose ECG data hiding technique for telemedicine applications. *Pattern Recognit Lett* 2019; 125: 463–73.
18. Goodrich MT. The mastermind attack on genomic data. In: 30th IEEE symposium on security and privacy. IEEE; May 17, 2009: 204–18; Oakland, CA.
19. Sweeney L. k-anonymity: a model for protecting privacy. *Int J Unc Fuzz Knowl Based Syst* 2002; 10 (05): 557–70.
20. Dwork C. Differential privacy. *Int Colloq Autom Lang Program* 2006; 4052 (d): 1–12.
21. Papadimitriou S, Li F, Kollios G, Yu PS. Time series compressibility and privacy. In: proceedings of the 33rd international conference on very large data bases; Sep 23, 2007: 459–70; Vienna, Austria.
22. Fan L, Xiong L. An adaptive approach to real-time aggregate monitoring with differential privacy. *IEEE Trans Knowl Data Eng* 2014; 26 (9): 2094–106.
23. Xiao X, Wang G, Gehrke J. Differential privacy via wavelet transforms. *IEEE Trans Knowl Data Eng* 2011; 23 (8): 1200–14.
24. Beaulieu-Jones BK, Wu ZS, Williams C, *et al.* Privacy-preserving generative deep neural networks support clinical data sharing. *Circ Cardiovasc Qual Outcomes* 2019; 12 (7): e005122.
25. Alvim M, Chatzikokolakis K, Palamidessi C, Pazii A. Local differential privacy on metric spaces: optimizing the trade-off with utility. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF). IEEE; Jul 9, 2018: 262–7; Oxford, UK.
26. Chatzikokolakis K, Andrés ME, Bordenabe NE, Palamidessi C. Broadening the scope of differential privacy using metrics. In: international symposium on privacy enhancing technologies symposium; Jul 10, 2013: 82–102; Springer, Berlin, Heidelberg.
27. Andrés ME, Bordenabe NE, Chatzikokolakis K, Palamidessi C. Geo-indistinguishability: differential privacy for location-based systems. In: proceedings of the 2013 ACM SIGSAC conference on computer & communications security. ACM; Nov 4, 2013: 901–14; Berlin, Germany.
28. Xiang Z, Ding B, He X, Zhou J. Linear and range counting under metric-based local differential privacy. In: 2020 IEEE international symposium on Information Theory (ISIT). IEEE; Jun 21, 2020: 908–13; Los Angeles, CA.
29. Fan L, Bonomi L. Time series sanitization with metric-based privacy. In: IEEE Big Data Congress; Jul 2, 2018: 264–7; San Francisco, CA.
30. He X, Machanavajjhala A, Ding B. Blowfish privacy: tuning privacy-utility trade-offs using policies. In: proceedings of the 2014 ACM SIGMOD international conference on management of data. ACM; Jun 18, 2014: 1447–58; Snowbird, UT.
31. Kifer D, Machanavajjhala A. Pufferfish: a framework for mathematical privacy definitions. *ACM Trans Database Syst* 2014; 39 (1): 1–36.
32. Thakurta AG, Vyrros AH, Vaishampayan US, *et al.* Emoji frequency detection and deep link frequency. United States patent US 9,705,908; July 2017.
33. Tang J, Korolova A, Bai X, Wang X, Wang X. Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. *arXiv e-prints* 2017:arXiv-1709.
34. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. *Nat Rev Genet* 2014; 15 (6): 409–21.
35. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. *PLoS One* 2011; 6 (12): e28071.
36. Fernandes AC, Cloete D, Broadbent MTM, *et al.* Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Med Inform Decis Mak* 2013; 13 (1): 1–14.
37. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. *BMJ* 2015; 350: h1139.
38. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. In: machine learning for healthcare conference. PMLR; Nov 6, 2017: 286–305; Boston, MA.
39. Bonomi L, Jiang X, Ohno-Machado L. Protecting patient privacy in survival analyses. *J Am Med Inform Assoc* 2020; 27 (3): 366–75.
40. De Montjoye Y-A, Hidalgo CA, Verleysen M, Blondel VD. Unique in the crowd: the privacy bounds of human mobility. *Sci Rep* 2013; 3: 1376.
41. Yan C, Zhang Z, Nyemba S, Malin BA. Generating electronic health records with multiple data types and constraints. In: AMIA annual symposium proceedings. Vol. 2020. American Medical Informatics Association; Nov 14, 2020: 1335; Virtual.
42. Keogh E, Ratanamahatana CA. Exact indexing of dynamic time warping. *Knowl Inf Syst* 2005; 7 (3): 358–86.
43. Liu Y, Yan C, Yin Z, *et al.* Biomedical research cohort membership disclosure on social media. In: AMIA annual symposium proceedings. Vol. 2019. American Medical Informatics Association; Nov 16, 2019: 607; Washington, DC.
44. Umar P, Akiti C, Squicciarini A, Rajtmajer S. Self-disclosure on Twitter during the COVID-19 pandemic: a network perspective In: joint European conference on machine learning and knowledge discovery in databases; Sep 13, 2021: 271–86; Bilbao, Spain.
45. MIT-BIH Supraventricular Arrhythmia Database. https://physionet.org/content/svdb/1.0.0/. Accessed September 1, 2021.
46. Olszewski RT. *Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data*. Pittsburgh, PA: Carnegie Mellon University; 2001.
47. Pimentel MAF, Johnson AEW, Charlton PH, *et al.* Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Trans Biomed Eng* 2017; 64 (8): 1914–23.
48. Goldberger AL, Amaral LAN, Glass L, *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 2000; 101 (23): e215–20.
49. Kate RJ. Using dynamic time warping distances as features for improved time series classification. *Data Min Knowl Disc* 2016; 30 (2): 283–312.
50. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; 9 (11): 2579–605.