applied sciences

MDPI

Article

Privacy and Utility of Private Synthetic Data for Medical Data Analyses

Arno Appenzeller [1,2,*], **Moritz Leitner** [1,2], **Patrick Philipp** [2], **Erik Krempel** [3] and **Jürgen Beyerer** [1,2]

[1] Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
[2] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, 76131 Karlsruhe, Germany
[3] Department of Computer Science and Mathematics, Hochschule München University of Applied Sciences, 80335 München, Germany
* Correspondence: arno.appenzeller@kit.edu

Abstract: The increasing availability and use of sensitive personal data raises a set of issues regarding the privacy of the individuals behind the data. These concerns become even more important when health data are processed, as are considered sensitive (according to most global regulations). Privacy-Enhancing Technologies (PETs) attempt to protect the privacy of individuals whilst preserving the utility of data. One of the most popular technologies recently is Differential Privacy (DP), which was used for the 2020 U.S. Census. Another trend is to combine synthetic data generators with DP to create so-called private synthetic data generators. The objective is to preserve statistical properties as accurately as possible, while the generated data should be as different as possible compared to the original data regarding private features. While these technologies seem promising, there is a gap between academic research on DP and synthetic data and the practical application and evaluation of these techniques for real-world use cases. In this paper, we evaluate three different private synthetic data generators (MWEM, DP-CTGAN, and PATE-CTGAN) on their use-case-specific privacy and utility. For the use case, continuous heart rate measurements from different individuals are analyzed. This work shows that private synthetic data generators have tremendous advantages over traditional techniques, but also require in-depth analysis depending on the use case. Furthermore, it can be seen that each technology has different strengths, so there is no clear winner. However, DP-CTGAN often performs slightly better than the other technologies, so it can be recommended for a continuous medical data use case.

Keywords: synthetic data generation; differential privacy; secondary use; medical data; private data processing; open source framework

check for updates

Citation: Appenzeller, A.; Leitner, M.; Philipp, P.; Krempel, E.; Beyerer, J. Privacy and Utility of Private Synthetic Data for Medical Data Analyses. *Appl. Sci.* **2022**, *12*, 12320. https://doi.org/10.3390/app122312320

Academic Editors: Petr Dzurenda, Sara Ricci and Jordi Castellà-Roca

Received: 2 November 2022
Accepted: 27 November 2022
Published: 1 December 2022

1. Introduction

In recent years, the digitization of the healthcare sector has been advancing more and more. A recent example is the long-awaited soft launch of the German electronic health record "Elektronische Patientenakte" (ePA) for all state-insured patients. The ePA and its corresponding infrastructure provides a platform for nearly every aspect of digital healthcare. The ePA itself stores the personal health data of a patient that arise during a patient's treatment [1]. The main goals of these data are to facilitate data sharing between different caregivers to improve patient care as well as to give patients better insight into their treatment. Another aspect of this growing amount of personal health data is that these can also be an important asset in medical research. Having access to digital, structured medical data promises tremendous benefits for various secondary use scenarios such as big data analysis in clinical research [2].

Besides those potential benefits, there are many open questions, especially regarding data privacy. The General Data Protection Regulation (GDPR) of the European Union considers medical data to be sensitive information, the processing of which is generally prohibited according to Article 9 (1). In practice, many medical studies use anonymization or pseudonymization to protect individual privacy. Nevertheless, as several incidents show, there is a risk of re-identification for an individual even with data that were assumed to be anonymous by the responsible parties. For example, Latanya Sweeney was able to uniquely identify the Massachusetts Governor William Weld in a purportedly anonymized data set from the Massachusetts Group Insurance Commission by linking it to a public voter registry [3]. To mitigate these risks of re-identification, a variety of Privacy-Enhancing Technologies (PETs) exist. More traditional techniques such as k-anonymity or l-diversity provide a conservative privacy quantification and can also help to tailor a data set in terms of privacy protection. More sophisticated technologies such as Differential Privacy (DP) provide a statistical measure that quantifies the privacy of the individuals contained in the dataset [4].

All those technologies involve a trade-off between privacy and utility. The example of DP shows that a dataset that provides a high privacy guarantee often lacks utility and cannot be used in practice [5]. With breakthroughs in machine learning due to improved algorithms and increases in computational power in recent years, the generation of synthetic data has attracted the interest of researchers as well as practitioners [6]. These techniques promise to create accurate and at the same time private data. This is because the raw data are only used for training and the result is then purely synthetic while preserving the statistical properties of the raw data. However, it remains to be seen whether such characteristics are not a privacy leak in themselves [7].

In this work, we investigate the feasibility of synthetic data to preserve privacy and utility for an exemplary medical dataset. Therefore, the use case of a data donation is introduced, where the heart rate data of individuals are collected. By generating private synthetic data in this scenario, a whole set of privacy issues could be solved, which could also lead to higher acceptance and participation in such data donations. We analyze the extent to which privacy is really protected with synthetic data and how high the utility of the data is compared to established technologies such as DP.
The contributions of this work are the following:

- Overview of existing technologies and implementations for generating private synthetic data;
- Introduction of a use case-specific privacy and utility metric for synthetic medical data;
- Evaluation of three approaches for private synthetic data generation in terms of privacy and utility.

Our article is structured as follows: Section 2 of this paper gives an overview of related work in the field. In Section 3, we summarize different PETs that can be considered for a medical data analysis use case. Such a use case is outlined in Section 4, while Section 5 describes our experiments with this use case. These experiments are evaluated and discussed in Sections 6 and 7. We then conclude the article in Section 8 and provide an outlook for future work.

2. Related Work

There is a multitude of papers that discuss private synthetic data and their potential applications. "A Privacy Mirage" by Stadler et al. [7] questions the promise that synthetic data provide perfect privacy protection. The authors conducted a quantitative assessment on relevant privacy concerns such as relinkability using an evaluation framework for synthetic datasets. The paper concludes that synthetic data suffer from the same privacy–utility trade-off as traditional anonymization techniques. While our work addresses a similar research question, we evaluate synthetic data in terms of a specific medical use case where particularly high data accuracy and correlation between different data points is important.

Bellovin et al. [8] gave an overview of synthetic datasets from a legal viewpoint. The authors compared the approach of generating synthetic data with traditional privacy measures and explained the different legal requirements for private data. However, the article notes that synthetic data can have significant benefits for sharing private data but must be done properly. In addition, while current laws permit synthetic data generation, they may not consider every aspect regarding the risks and benefits of this approach.

The article "Synthetic data in machine learning for medicine and healthcare" gives a broad overview about the topic of synthetic data for medical use cases [9]. The authors ask what risks or benefits synthetic data generation implies for patients. One of the main considerations is that generation through Generative Adversarial Network (GAN)s not only bears the risk of creating deepfakes used for abuse, but can also help create anonymized data. Besides the potential benefits, the authors also urged for regulatory standards and metrics that should also address information leakage which might occur through synthetic data generation. While this commentary provides a rather general overview, the measurement of potential information leakage is a key issue that our work aims to address.

Bowen and Snoke [10] published a comparative study on differential private synthetic data algorithms which were part of the NIST PSCR Challenge focused on synthetic private data. In this study, the authors focused on accuracy and usability for data providers. Like our work, the paper presents two custom utility metrics. They examined a range of different values for the privacy budget ϵ and visualized the changes in the utility of the data as ϵ increases. The authors conclude that there is no one-size-fits-all algorithm. Instead, potential users should determine which metrics are appropriate for their use case to evaluate their choice of algorithm. While this paper focuses on the utility and usability of synthetic data, we try to consider and evaluate the real-world privacy implications of such techniques in our work.

The paper "Privacy Preserving Synthetic Data Release Using Deep Learning" by Abay et al. [6] compares existing technologies with different utility metrics and provides a novel approach to generate differential private datasets. The introduced algorithm is a generative autoencoder which partitions the input data into groups, in which a private autoencoder (which adds noise to the gradients using a DP mechanism) learns the structure of the group and tries to accurately simulate this data. This approach was benchmarked to other techniques, such as PrivBayes, by using traditional machine learning metrics. While there was no clear winner, the authors concluded that their approach performs well in various situations and yields robust results. In the benchmark section, Abay et al. did not provide a real-world use case as we do. Moreover, they also focus on utility metrics and not the actual privacy impact besides the ϵ value.

Another publication by Rosenblatt et al. [11] also evaluated differential private synthetic data generation. The authors benchmarked four different GAN-based algorithms for generating synthetic data against common machine learning metrics. In addition, the authors introduced an ensemble method called QUAIL, which can be embedded in other DP-GANs and enables an even distribution of the privacy budget. The benchmark showed that the combination with QUAIL outperforms the baseline DP generators. The result is like others, with no clear winner, but PATE-CTGAN often has a better performance due to higher utility and statistical similarity when using higher ϵ values. Unlike our work, Rosenblatt et al. measured precision and utility on various datasets without a concrete domain-specific use case. Additionally, the focus is on assessing the utility rather than the degree of privacy of the data.

A more traditional concept was pursued for the DataSynthesizer project by Ping et al. [12]. Here, a three-component workflow is used to create synthetic data based on tabular input data. At first, the `DataDescriber` reads the structure, correlation, and distribution of the data. From this, the `DataGenerator` creates a synthetic dataset using the data points. The `ModelInspector` analyzes this dataset and provides parameters to iteratively refine the data quality of the synthetic dataset. In contrast to other technologies, `DataSynthesizer` does not use a GAN approach or a DP mechanism. Although the concept appears to be interesting, we exclude it from our work as it seems too far from the scope of the technologies used here.

A paper that proves the real-world feasibility of synthetic data is "Differentially Private Medical Texts Generation Using Generative Neural Networks" [13]. Al Aziz et al. used a differentially private training method in combination with the Generative Pre-trained Transformer 2 (GPT-2) language generation model, where Gaussian noise is added on the weights in every iteration of the training process. To evaluate their approach, the authors used several metrics to compare the generated data to the reference data set. From a utility perspective, the approach received good scores, but the authors remarked that they did not consult medical experts to verify the plausibility of the medical data. In addition, they also noted that they did not de-identify the data or include anything to prevent the model from learning identifying data. This also underlines our research question to pay more attention to use case-specific attacks on privacy. Nevertheless, the paper shows that synthetic data are a promising tool for the medical domain.

In another work by Carlini et al. [14], the issue of data correlations and preserved patterns that could seriously undermine privacy is examined. An exposure metric is introduced, which measures the characteristics of the dataset. Using this metric, the problem of unintended memorization is demonstrated on some sample data sets. The authors recommend strategies such as regularization, sanitization, as well as DP to mitigate those issues. However, these strategies cannot be applied in general, and our approach focuses on techniques that already use DP.

The just presented small share of all papers that address the topic of synthetic private data generation shows the broad interest of the community in this topic. However, we found no work that conducts an evaluation like ours by choosing a real-world use case, defining a scenario-specific privacy definition, and benchmarking different technologies for that use case.

3. Data Privatization Techniques

As a prerequisite for our experiments, the definitions of several PETs will be recalled in this section. At first, more traditional techniques such as k-anonymity and l-diversity are described. Differential privacy is also presented, as it is used for comparison with synthetic private data generation, which is also outlined in this section. Following the k-anonymity and l-diversity data model, personal data can be categorized into at least three types [3,15]:

- Identifier: A data point that can be used to uniquely identify a person, e.g., a name or a personal ID.
- Quasi-Identifier: A data point that must be combined with other quasi-identifiers to allow inferences, e.g., zip code, age, gender.
- Sensitive attribute: A data point that holds some sensitive information about a person, e.g., a concrete illness.

Please keep in mind that these types are not mutually exclusive. A concrete illness can be sensitive information and a quasi-identifier at the same time. Potentially, this could be a concern with genetic data analyses in the near future.

3.1. Traditional Techniques

A common approach for protecting the privacy of individuals contained in a dataset is k-anonymity, which was first introduced by Sweeney [3]. The main principle of k-anonymity is to first remove all identifiers and then transform the dataset so that there are groups of k entries with the same quasi-identifier. Records with matching values on quasi-identifying attributes form an equivalence class, preventing a sensitive attribute from being linked to a single person [16]. This can be achieved by the generalization and suppression of the (quasi-)identifying attributes or by the addition of fake data. When using generalization, the value of a quasi-identifier is replaced with a more general super category. For example, a certain age will be replaced with a suitable age range. Suppression can be used for attributes where there is no good way to generalize the values. Then, the value can be suppressed, which in general means the removal of the value [3].

k-anonymity has non-negligible weaknesses. For example, an attacker with background knowledge can assign a specific probability to each sensitive attribute within an equivalence class, which can significantly compromise an individual's privacy. To address this weakness, the definition of *l*-diversity was evolved [15]. This is based on *k*-anonymity and uses the same suppression and generalization techniques but enforces a stronger privacy guarantee. *l*-diversity necessitates that there are at least *l* distinct sensitive attributes per equivalence class. While this protects from some possible attacks in *k*-anonymity, it still has many pitfalls. If an attacker learns that their target has either prostate cancer or colon cancer, then highly sensitive information has already been leaked. With knowledge of the target's gender, this information might even become more accurate. There are also more sophisticated definitions such as *t*-closeness or δ-presence that try to remedy these shortcomings but have their own weaknesses. Since their development and use was mostly discontinued after DP was released, they are not discussed in this paper [17,18].

3.2. Differential Privacy

The definition of Differential Privacy (DP) was introduced by Cynthia Dwork in 2006 [19]. The aim was to establish a privacy guarantee that is independent of any background knowledge. One of the main principles is that it should make no difference for an individual's privacy whether their data are included in a dataset. This property is also reflected in the definition of DP.

Definition 1 (ϵ-Differential Privacy)**.** *A randomized algorithm $\mathcal{K}$ is ϵ-differentially private if, for all $\mathcal{S} \subseteq Range(\mathcal{K})$ and for all databases D_1 and D_2, which differ by at most one entry, the following holds [20]:*

$$Pr[\mathcal{K}(D_1) \in \mathcal{S}] \leq e^{\epsilon} \times Pr[\mathcal{K}(D_2) \in \mathcal{S}].$$

The choice of ϵ, which is also called privacy budget, is non-trivial. Although it is obvious that a small ϵ leads to higher privacy, there are a lot of different values that are recommended or used in the literature. The choice also depends on the use case and which privacy–utility trade-off is acceptable. In addition, there are different ways to fulfill the definition. One common way is to use the Laplacian-distributed noise, while other methods work with randomized responses. Furthermore, there are different types of DP. In the central model, the data are stored in a central database. This database contains the raw data, so individuals must trust it. DP algorithms are applied when data or aggregated results are requested. The local variant of DP applies the DP mechanism before the data are shared. Thus, individuals are not required to trust a central database and instead privatize the data before sharing it.

3.3. Synthetic Private Data

There were very early concepts of synthetic private data based on the statistical properties of the data, for example, by Rubin [21]. More recently, and with the advances in machine learning, the idea of synthetic private data came back into the focus of research. One method is to use variational autoencoders (VAEs) to generate new data based on a training dataset [22]. To preserve privacy, in this case, the learning accuracy can be limited to avoid overfitting that could reveal private information from the training data. However, the current trend is usually based on generative models, in particular, on so-called GANs, where two networks are in a game to create synthetic data that are indistinguishable from the training data [23]. In most settings, there is one network that uses training data to generate new synthetic data based on the input, and another network judges whether the data are realistic or resemble artificial data. This process is often used for image-based synthetic data generation (known as deep fakes) but can also be used for text-based or tabular data. For the GAN-based approach, the idea of using DP in the data generation process came up to enable a *real* privacy guarantee.

Given this foundation, we used three popular private synthetic data algorithms that are implemented in the SmartNoise Synthesizers framework [24]. Among them is the histogram-based approach Multiplicative Weights Exponential Mechanism (MWEM) by Hardt et al. [25]. The algorithm maintains an approximation of the data distribution. In each iteration, the worst approximated query is selected using the exponential mechanism. Subsequently, the accuracy is increased according to the multiplicative weights update rule. Through this process, a dataset can be generated that mimics the input distribution.

DP-CTGAN, first introduced by Rosenblatt et al. [11], uses the deep learning-based CTGAN model proposed by Xu et al. [26] as its foundation. CTGAN generates tabular data using a generator network that can incorporate conditional distributions to avoid class imbalance problems. These generated data are then passed to a discriminator to judge the quality of the synthetic data compared to the original data. In this step, DP noise is added to privatize the results.

Finally, PATE-CTGAN, also first outlined by Rosenblatt et al. [11], likewise uses CTGAN but employs the private teacher ensemble (PATE) [27] method to split the input data into partitions. These partitioned data are used to train different teacher networks that predict labels for the data. DP noise is also added in this process to privatize the data. Then, the student networks learn with the teacher's privatized labels, resulting in synthetic data.

4. Medical Use Case

In this section, we outline the stringent legal requirements for protecting medical data. Furthermore, the scenario of this paper, a medical data donation of lifestyle fitness data, is described. Additionally, challenges regarding privacy are discussed and an attacker model is introduced.

4.1. Legal Requirements

Medical data processing requires the highest standard of data protection. The European GDPR considers medical data as sensitive data. Hence, processing is only allowed if one of few exemptions can be applied, e.g., when data processing is required for a medical treatment according to GDPR Article 9 (2)(a) and (h). For the secondary use of medical data, such as research purposes, the most common exception criterion is an explicit declaration of consent by the affected person. Nevertheless, strong privacy protection measures are required.

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) defines privacy rules for the usage of personal health data [28]. Like the GDPR, it follows the principle that patients must give their consent before their data are processed. Additionally, there are exclusions where processing is mandatory, such as for payment purposes or when the data are needed for treatment. HIPAA also regulates the secondary use of personal health data and provides guidance on the de-identification of data (§ 164.514). Included in this guidance is a list of 18 identifiers which should be removed so that the data can be considered de-identified. Such identifiers are the name of the subject, social security number, etc. However, HIPAA's de-identification guidelines show another issue regarding the privacy of personal medical data. While HIPAA considers data as de-identified when all specified identifiers have been removed, it does not address the issue of re-identification. The GDPR contains similar provisions for anonymized data without stating specific identifiers, but also does not consider the issue of re-identification. Potential re-identification can occur in many scenarios, especially if an attacker has background knowledge that can be used to link quasi-identifiers (e.g., zip code, year of birth) between different datasets. As such, an individual's identity can be revealed, potentially compromising their privacy. There are many studies that show the risk of re-identification in de facto anonymized datasets [29–32]. This has led many legal scholars to demand that the state of the art must be considered in the case of anonymization. For this reason, PETs like the ones of Section 3 are used. In this article, synthetic differential private data are utilized to mitigate the re-identification risk.

4.2. Secondary Use Scenario

In this work, a scenario of secondary use of medical data are considered. Within this scenario, several individuals donate their data to a research project which aims to analyze the data for certain properties. Fitness trackers and wearables are becoming increasingly popular, as users of these devices generate a large amount of health data that could also be of interest to researchers. The most common data collected by such trackers are the heart rate and step count. Continuous heart rate monitoring can provide deep insights into an individual's health. One property that can be identified by tracking the heart rate is the presence of what is known as tachycardia [33]. Tachycardia can be diagnosed if the heart rate exceeds 100 beats per minute (BPM). In most cases, this can be explained through physical activity. However, tachycardia at rest can be an indicator of more serious health conditions. In our scenario, wearable users donate their heart frequency data, which are then analyzed for tachycardia. Although this seems like a superficial analysis, this paper aims for a real-world use case where accurate medical data are needed.

To simulate this scenario, a dataset of crowdsourced data from Fitbit health trackers is used [34] that was created by the RTI International Institute. The data were collected through a survey on Amazon Mechanical Turk. This dataset contains continuous heart rate measurements obtained by Fitbit wearables from 30 participants over a one-month period. In detail, the approximately 15 million data points including heart rate measurements (5- or 10-s interval), step counts (60-s interval) and the duration of sleep. Each entry of the dataset has a timestamp and a participant ID. For privacy reasons, there is no additional information or demographic data about the participants. It should be noted that only the heart frequency measurements are used in this work (`heartrate_seconds_merged.csv`). Figure 1 shows an exemplary plot of one of the heart rate measurements of the dataset.

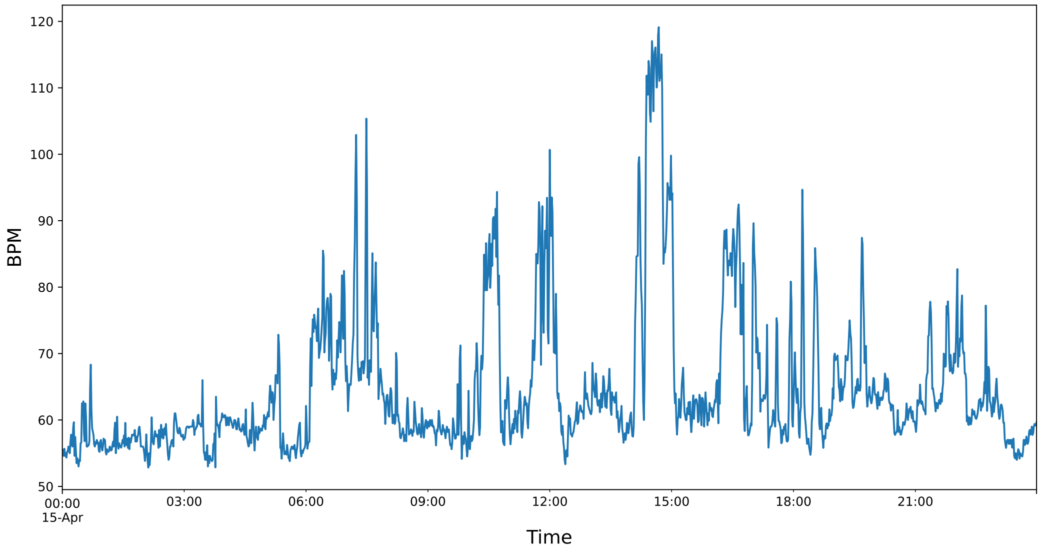


Figure 1. Heart rate of an individual from the Fitbit dataset for one day.

These measurements follow a typical scheme: There are some phases with a higher heart rate (during daytime) and lower heart rate at night. Furthermore, there are some peaks that may indicate exercise or other physical activity. Even such a small subsample of data shows that it is possible to detect the activity patterns of individuals which might not be obvious when sharing these data.

4.3. Privacy Challenges

As mentioned before, the data that are used here can reveal privacy-sensitive information about an individual. For example, the time course of the raw data reveals a person's sleep or activity time. The usual method to mitigate such risks is to use DP, where noise is added to all data points to make them unlinkable to the raw data. Typical applications for DP include the calculation of average or median values. The concept of DP assumes that the noise added to single data points is canceled out over the entire dataset. This should allow an accurate analysis. For the tachycardia use case described in Section 4.2, average or median values alone are not sufficient. An accurate estimate of the minutes with a heart rate over 100 BPM is required. In an interactive DP setting, this is not easily possible because the raw data must be transformed into one-minute intervals. Therefore, a static data release (or the so-called non-interactive setting) is much better suited for such an analysis [35]. In the following section, we will conduct a plain DP release and use synthetic data technologies to create a private dataset with the same number of entries and individuals. The idea of using synthetic private data in this context is that the statistical properties of the data (e.g., the Tachycardia minutes) are preserved, but the activity patterns are disguised. The typical measure for privacy when using DP is ϵ. While ϵ quantifies a mathematical property, choosing an appropriate value to guarantee privacy for a real-world use case is nontrivial to impossible.

To motivate the subsequent analyses of this paper, an attacker model is introduced. This attacker has the ability to recognize activity patterns from the heart rate data. The attacker further has the necessary medical knowledge and can understand the heart rate data. Additionally, the attacker has background knowledge about its victim, such background knowledge can be the typical wake-up time of the victim. In combination with the heart rate graph, the attacker can use this knowledge to re-identify the victim and link it to the *anonymized* heart rate data. These data can then be used to gain additional information about or, in the worst case, to blackmail the victim.

Definition 2 (Re-identification attacker $\mathcal{A}$)**.** *A re-identification attacker $\mathcal{A}$ combines its medical and victim background knowledge to re-identify a victim using anonymized heart rate data. $\mathcal{A}$ explicitly uses activity patterns of the heart rate graph for re-identification. Potential background knowledge can be the typical sleep or wake-up time or more precise knowledge about daily routines such as workout times. The attacker is successful if it was possible to link a certain individual to data from a heart rate dataset containing no direct identifiers.*

To measure privacy besides the statistical ϵ, we define a domain-specific metric called *Tachycardia Privacy (TP)*. As described previously, changes in the heart rate curve are considered as private information because they reveal activity patterns or daily routines. The goal of a private dataset, which was generated using DP or synthetically, should be to have a heart rate curve with a gradient that differs as much as possible from the gradient of the real curve. TP is hence meant to quantify how likely it is that an attacker could use the privatized data to re-identify an individual if they have knowledge of an individual's daily activities. The principle is to measure the similarity of the derivative of the heart rate curve at each point in time for the raw data and the privatized data, respectively. For this purpose, functions $f'_{raw}(t)$ and $f'_{private}(t)$ are needed, which represent the derivative of the heart rate curve at a point in time $t \in T$. T is the normalized time for all heart rate curves, where the earliest time is 1 and the latest t_n. In this work, one day is considered as a time period, thus $1 \leq t \leq 1440 = t_n \land t \in \mathbb{N}$ (t can be any minute of one day). In addition, we use the so-called *Privacy Impact Factor (PIF)* in the TP to weight the impact of the relative difference between the two derivatives.

Definition 3 (Tachycardia Privacy (TP))**.**

$$TP = \frac{\sum_{t \in T} min\left(\frac{|f'_{raw}(t) - f'_{private}(t)|}{f'_{raw}(t)}, PIF\right)}{|T| \times PIF}$$

The core idea of Tachycardia Privacy (TP) is to measure the difference between the derivatives of the raw and privatized heart frequency curves. From a privacy perspective, a maximal different derivative leads to a maximal different heart rate curve. With this, our previously described attacker has no way to make use of their background knowledge to re-identify an individual. To accomplish this, we use the absolute difference between the gradient of the raw and privatized heart frequency and put it in to relation to the raw gradient. The Privacy Impact Factor (PIF) defines the maximum difference between the raw and synthetic curve which is considered. A $PIF = 1$ *means that every difference larger than 100% will not have a greater impact on the value because of the minimum function. For our experiments, we saw that a* $PIF = 3$ *gives the best range of values for the privacy results. With this, relative differences of up to 3x are considered in the calculation. We do this calculation now for all time points in our heart frequency series. To arrive at the final TP between 0 and 1, the summed value is divided through the number of entries times PIF. In summary, this means the closer the TP is to one, the better the privatization, because the courses of the synthetic and raw heart rates differ more, making re-identification less likely.*

While privacy is an important factor and should be a fundamental principle in the processing of any health data, the data must also be usable. For this, a definition of utility is required, which we call *Tachycardia Utility (TU)*. To measure utility for the tachycardia use case, the number of minutes with a heart rate above 100 BPM (hereafter referred to as *tachycardia minutes*) must be compared with the tachycardia minutes of the generated data.

Definition 4 (Tachycardia Utility (TU))**.**

$$TU = 1 - min(1, \frac{|private\ minutes - raw\ minutes|}{raw\ minutes})$$

The Tachycardia Utility (TU) measures the deviation between the tachycardia minutes of the privatized and the raw data. To that end, we use the absolute difference between the private and raw minutes and divide it by the raw value to put this difference in relation to the size of the raw value. We also define that a difference of more than 100% is the worst possible and that such a privatized value has no more utility. Therefore, we limit the result to 1 using the minimum function. A TU of 1 is considered a perfect utility, since the number of minutes is the same. The closer the value is to 0*, the worse the utility is.*

Finally, these metrics allow us to measure the vulnerability and utility of the privatized data in the forthcoming experiments.

5. Experiments

In the following, the setup for the experiments performed in this paper is described. Furthermore, the plain DP experiments are outlined as a preliminary for the evaluation in the next section.

5.1. General Setup

SmartNoise is developed by the OpenDP initiative, which includes Microsoft as well as the Institute for Quantitative Social Science (IQSS) and the School of Engineering and Applied Sciences (SEAS) at Harvard University [24]. A broad application of the tools in research, industry, and by governmental institutions is the declared goal of the project participants. The open source framework is one of the most popular in the field of DP and comprises several components, including SmartNoise SQL and SmartNoise Synthesizers (https://github.com/opendp/smartnoise-sdk/tree/main/synth (accessed on 18 August 2022)). The former is used in the interactive scenario to privatize SQL queries against, for example, PostgreSQL databases or Pandas data frames. As the name suggests, the latter allows generating synthetic data with three different synthesizers (MWEM, DP-CTGAN, and PATE-CTGAN), which we already described in Section 3.3. The models of DP-CTGAN and PATE-CTGAN are trained with PyTorch, and for the discriminator of DP-CTGAN, the

Opacus library is used. We increased the default value of `max_bin_count` of MWEM to 1440 (number of minutes in a day) for our experiments, which turned out to be the best value for time series data such as those in our use case after a few trials. For PATE-CTGAN and DP-CTGAN, the default parameters were used.

We also made some adjustments to the Fitbit data set (`heartrate_seconds_merged.csv`) for our scenario. First, we only considered four individuals and shortened the examination time to one day, shrinking the dataset from nearly 2.5 million entries to approximately 35,000. In addition, the timestamps were changed to one-minute granularity and converted to the number of minutes since the beginning of Unix time to simplify later evaluations. After the entries were randomly sorted, this adjusted dataset (`tachycardia.csv`, hereafter referred to as raw data) finally served as the input for the synthetic data generators. The first five entries can be seen in Table 1.

Table 1. Excerpt from the heart rate dataset which was used as the input for the synthesizers.

Id	Rate	Unix_Min
2	94	24346062
1	73	24344778
1	69	24345716
2	61	24345101
2	70	24345613

Before calculating the metrics, both the synthetic and raw data were downsampled into one-minute bins by taking the mean to smooth out strong outliers and facilitate comparability. To determine the TP metric, the raw data as well as the generated data were first interpolated with cubic splines. Subsequently, the derivatives of the cubic splines were evaluated for each minute according to the previously presented definition of the metric. Afterwards, these values were added up and divided by the number of minutes multiplied by the PIF.

To calculate the TU, we simply counted the number of tachycardia minutes in both the raw and synthetic datasets. For all experiments, 20 runs were performed. The computations were executed on a machine with 2x Intel Xeon 4210 2.2 GHz CPUs, 8x NVIDIA GeForce RTX 2080 TI GPUs and 144 GiB RAM running CentOS 7 and Python 3.8.12.

5.2. Plain DP

To make a solid comparison, we also evaluated our metrics using plain DP technology. For this, we used the Laplace mechanism of the SmartNoise framework and applied it to all the heart frequency values. This should resemble a non-interactive DP data release. Such a data release gives data processors more possibilities to work with the data but comes with some privacy trade-offs. Figure 2 shows the Tachycardia Privacy (TP) and Tachycardia Utility (TU) for this setting.

It can be observed that the TP is low regardless of the choice of ϵ. Moreover, the range between the different runs is rather small (from almost 0 to 1.4). For the TU, it can be noted that regardless of the choice of ϵ, the values are close to 1. Furthermore, the values only scatter between 0.84 and 1. While this is generally good for the mechanism, the TP is not satisfactory. However, this phenomenon can be explained by the way plain DP was applied. In contrast to the synthetic data generation, plain DP does not generate a whole new dataset. Instead, noise is drawn form a Laplace distribution at each point of time to perturb the heart frequency measurements. For this purpose, the Laplace mechanism implemented in SmartNoise was used. Since even a small ϵ would not cause a complete outlier, the overall shape of the heart rate curve is preserved, resulting in a weak TP. While Figure 2 shows only one individual, this pattern can be observed for the other participants as well.

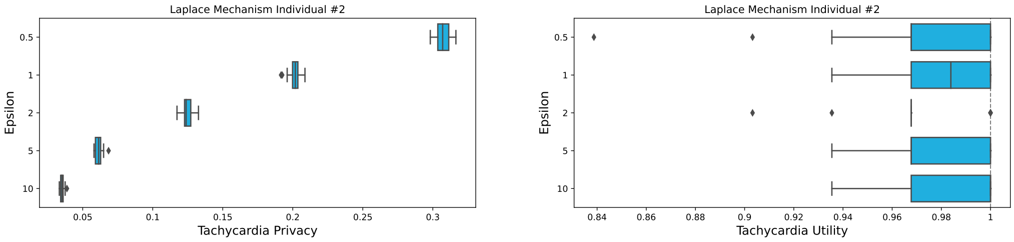


Figure 2. Box plots showing the Tachycardia Privacy (TP) and Tachycardia Utility (TU) for one individual using plain DP with Laplacian-distributed noise at 20 runs using a $PIF = 3$ for TP.

Synthetic generators, on the other hand, create an entirely new dataset, so they are a completely different approach. Therefore, plain DP technologies are not considered for further comparison. We conclude that while they provide good utility for each analyzed ϵ, they can potentially be vulnerable to an attacker like the one from Definition 2. Nevertheless, such issues can be solved by using DP in the interactive scenario and not releasing the complete dataset.

6. Evaluation

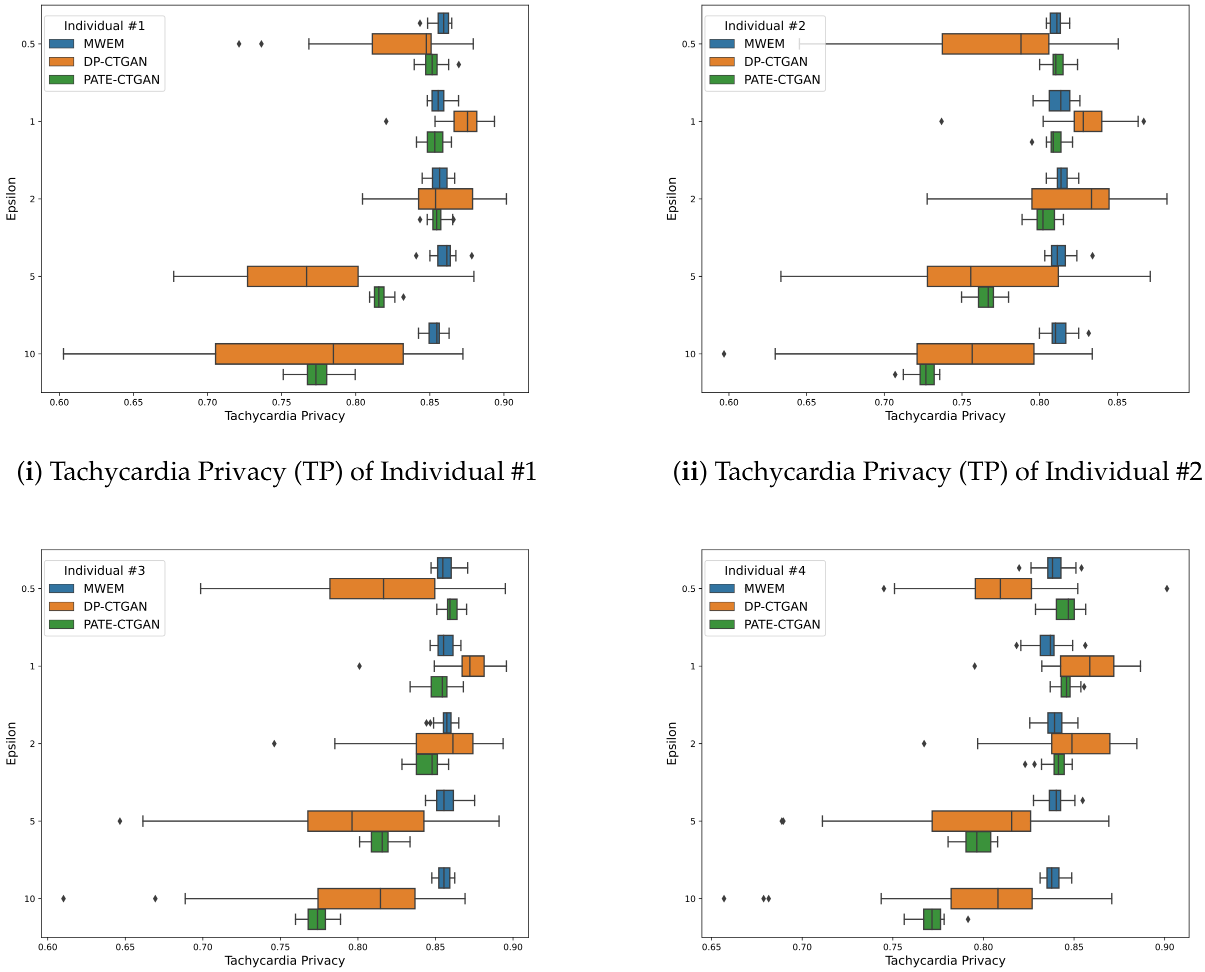
The evaluation of the three private synthetic data generators is presented in the upcoming section. At first, the generated data will be evaluated for privacy. After this, the utility will be assessed, and a ratio is calculated to find the sweet spot for the privacy utility trade-off. Finally, the performance measurements are shown as a runtime evaluation.

It should be noted that the individual with ID 4 is an outlier compared to others in this evaluation. Data for this participant is only available starting at 9 a.m., while the other participants provided data for the entire day. In addition, this individual has nearly twice as many tachycardia minutes throughout the day as the others.

6.1. Privacy

To evaluate privacy, we measured TP for the four individuals. Figure 3 shows the results of this measurement. Based on our expectation, TP should decrease with increasing ϵ as weaker and weaker noise is added in the generation process. It can be observed that the median values of DP-CTGAN and PATE-CTGAN in the box plots decrease with increasing ϵ, so the generated data become more and more accurate. With MWEM, on the other hand, the ϵ parameter seems to have at most a minor influence on the achieved TP.

For individual 1 in Figure 3i, the highest TP value is reached with DP-CTGAN and $\epsilon = 2$. The TP value is approx. 0.9 but should be viewed as an outlier. Considering only the median values, DP-CTGAN yields the highest value of approx. 0.87 at $\epsilon = 1$. The smallest median TP occurs for $\epsilon = 5$ when using DP-CTGAN, but the outer range with $\epsilon = 10$ includes smaller values. Individual 2, which is illustrated in Figure 3ii, exhibits the phenomenon that the median TP peaks at $\epsilon = 2$ and decreases thereafter as expected. Here, the smallest median value is obtained in the case of PATE-CTGAN and the highest ϵ. The participant with ID 3 in Figure 3iii shows similar results as the first two. As mentioned before, the raw data of individual with ID 4 differ from the other individuals. This also impacts the TP metric which can be seen in Figure 3iv. The number of outliers for all generators is much higher compared to the other individuals.



(**i**) Tachycardia Privacy (TP) of Individual #1

(**ii**) Tachycardia Privacy (TP) of Individual #2

(**iii**) Tachycardia Privacy (TP) of Individual #3

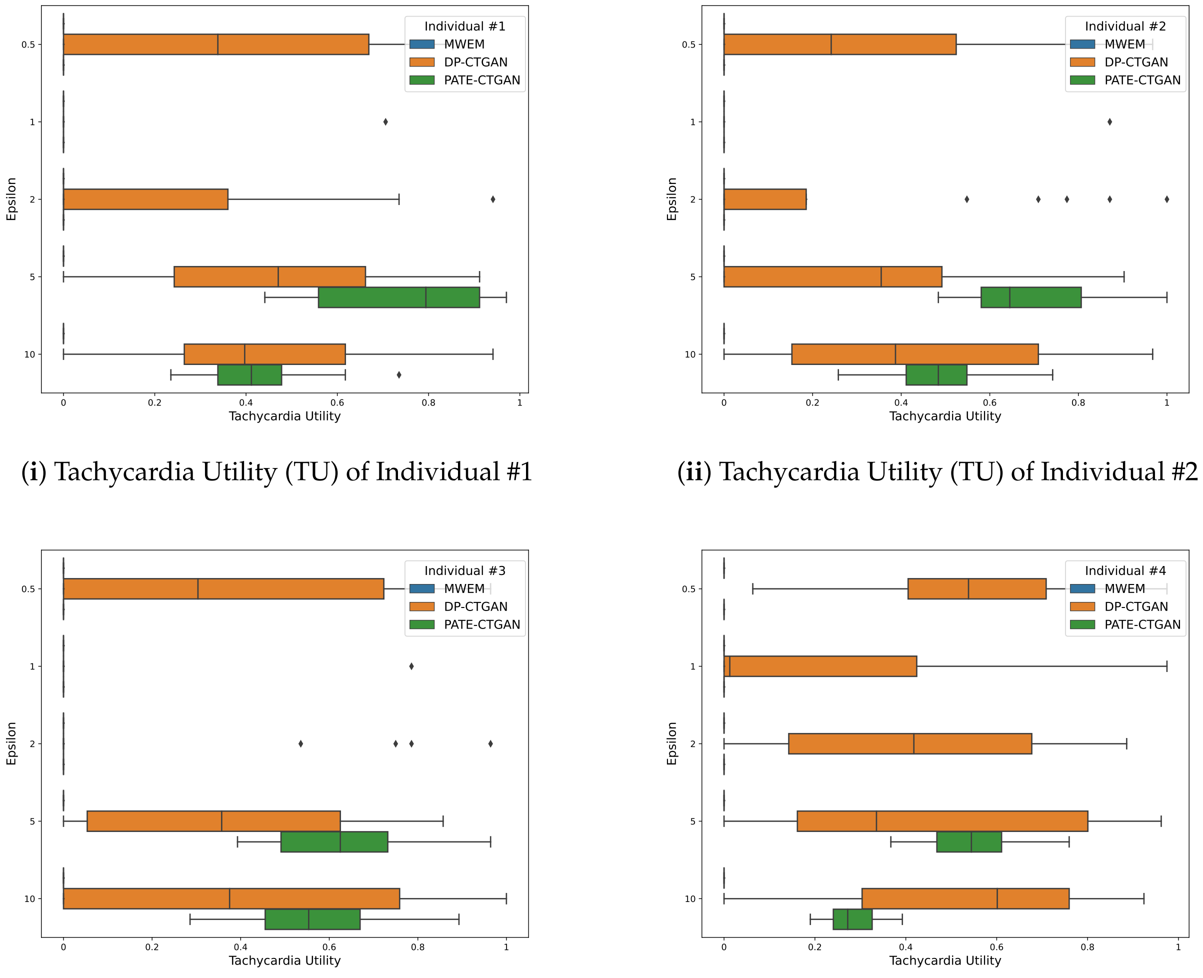
(**iv**) Tachycardia Privacy (TP) of Individual #4

Figure 3. Box plots showing the Tachycardia Privacy (TP) for four individuals using MWEM, DP-CTGAN, and PATE-CTGAN at 20 runs.

In summary, PATE-CTGAN shows the expected downward trend, which is also not consistent overall ϵ values. Additionally, a similar trend for DP-CTGAN can be observed but with a growing dispersion, the larger ϵ gets.

6.2. Utility

As for privacy, we measure TU as the utility metric for the four individuals. Figure 4 shows the results for this metric. In addition to this overview, Figure 5 provides a scatter plot for continuous ϵ values. To maintain readability, there is only one scatter plot for one individual. Nevertheless, the overall structure looks the same for the remaining participants. We anticipate that TU converges to the optimal value of 1 by using higher epsilon values for generating synthetic data, since less noise is added in the step of the mechanism employed.



(**i**) Tachycardia Utility (TU) of Individual #1

(**ii**) Tachycardia Utility (TU) of Individual #2

(**iii**) Tachycardia Utility (TU) of Individual #3

(**iv**) Tachycardia Utility (TU) of Individual #4

Figure 4. Box plots showing the Tachycardia Utility (TU) for four individuals using MWEM, DP-CTGAN, and PATE-CTGAN at 20 runs.

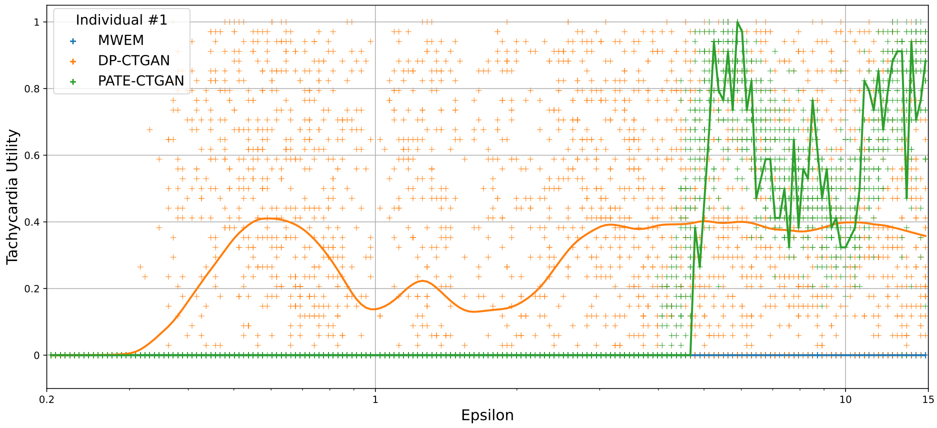


Figure 5. Scatter plot illustrating the continuous evaluation of the Tachycardia Utility (TU) for an individual using MWEM, **DP-CTGAN,!** and PATE-CTGAN at 20 runs.

Overall, the TU for MWEM is out of competition for all individuals. It remains at 0 for any ϵ and shows no improvement for larger ϵ values. This is indicated by the first black line above the values of DP-CTGAN in orange. In addition, the observation can be confirmed through the scatter plot in Figure 5, in which MWEM is a straight line. Moreover, there is no dispersion at all compared to the other technologies. For all individuals, it can be seen that PATE-CTGAN shows an improvement trend with a larger ϵ. The TU is around 0 for small ϵ values, and only starts to show usable values for $\epsilon = 5$. At $\epsilon = 5$, they are closely around 0.5 and 0.8. However, the largest $\epsilon = 10$ reaches significantly worse TU scores than with $\epsilon = 5$, indicating a saturation trend in terms of utility. This can be also seen in more detail with the continuous ϵ values of the scatter plot in Figure 5. The scatter plot indicates that there are two optimal ϵ values around 6 and 11 and that the TU decreases between these optima. In addition, there is a larger dispersion of the TU values in this phase. DP-CTGAN also shows the expected upward trend for TU for all individuals. However, there is a local optimum around $\epsilon = 0.6$ beyond which the utility declines. This can be seen in the scatter plot from Figure 5. After the TU moves away from the optimal value, the values slowly converge to 0.4 afterwards. This is consistent with the expectation that a higher ϵ leads to a saturated TU. It remains to be noted that DP-CTGAN has the highest dispersion of all technologies regardless of ϵ. The scatter plot confirms this observation. Participant number 4, shown in Figure 4iv, again has its typical outlier role with clearly deviating TU values. However, the trend observed with the other individuals for PATE-CTGAN remains the same.

6.3. Privacy Utility Trade-Off

Figure 6 visualizes the influence of ϵ on the so-called privacy–utility trade-off, which we also refer to as the privacy–utility Ratio. In an optimal setting, the choice of ϵ should maximize TU and TP. To calculate the plot, we added both values and divided the result by 2. Thus, the optimal case would be an ϵ where the ratio is close to one.

$$\text{Privacy-Utility Ratio} = \frac{TU + TP}{2}$$

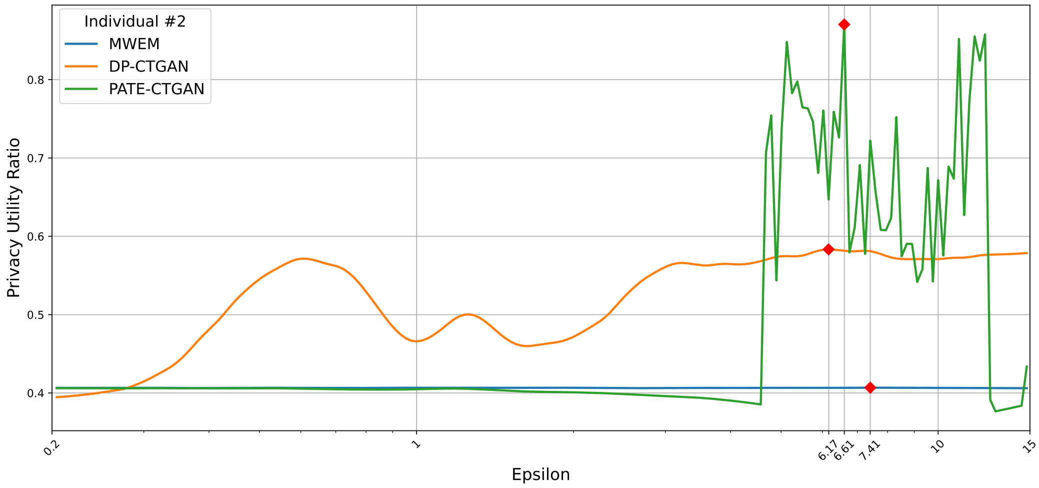


Figure 6. Plot of the privacy–utility ratio for an individual using MWEM, DP-CTGAN, and PATE-CTGAN at 20 runs. The red diamonds show the ϵ value at which the best trade-off between privacy and utility for each technique is achieved.

In Figure 6, the red markers indicate the optimal ϵ choice. For all three techniques, the optimal ϵ is between 6 and 7.5, after which the loss of privacy no longer not outweighs the potential utility benefit. For the other individuals, the results are similar, but with different, sometimes much smaller optimal ϵ values. However, it should be noted that in certain cases, such a ratio should also weight the values of TU and TP. Otherwise, for example, a perfect utility with no privacy could lead to a local maximum. Since such a weighting would be a very specific choice and the ratio worked well for our data, we excluded the weights.

6.4. Runtime

We anticipate the runtime for deep learning-based approaches (DP-CTGAN and PATE-CTGAN) to increase with growing ϵ, as a more complex model needs to be trained over an increasing number of epochs. In contrast, we expect that for MWEM, the runtime will remain relatively constant for different ϵ values, as it is a histogram-based concept. MWEM, unlike the aforementioned, also operates on the CPU and not on the GPU. The experiment results in Figure 7 show that our expectations are largely confirmed. The execution times for DP-CTGAN and PATE-CTGAN increase exponentially with rising ϵ, with a plateau forming for DP-CTGAN from approximately seven onwards. Furthermore, the runtime of MWEM varies only slightly around 350 s.

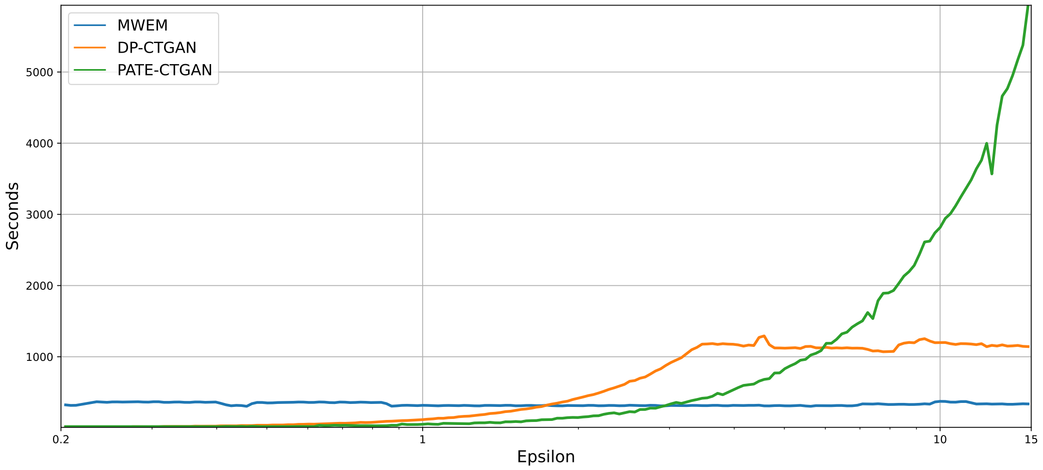


Figure 7. Runtime measurements for MWEM, DP-CTGAN and PATE-CTGAN.

7. Discussion

In this section, we discuss the results of the evaluation and state the limitations of this work.

7.1. Evaluation Results

At first, we would like to note at this point that we also carried out the experiments with Gretel Synthetics (https://github.com/gretelai/gretel-synthetics (accessed on 29 July 2022)) initially. We did not pursue this framework further after it became clear that it was not possible to choose a specific privacy budget ϵ. Instead, Gretel outputs the ϵ value only after a model with certain hyperparameters was trained. Furthermore, the ϵ values were in the three-digit range, which would not have permitted a meaningful comparison with SmartNoise.

One reason that the MWEM approach performs worse than the other technologies on both the TU and TP metrics could be that the concept of MWEM was not designed for time series or continuous values. MWEM should rather be used for categorical data [36]. For our experiments, we increased the default `max_bin_count` of SmartNoise to 1440 (number of minutes in a day), otherwise the results were even worse.

In terms of privacy, the influence of the increasing ϵ on our TP metric seems to be limited to a specific range. This is consistent with the observations from our plain DP experiments in Section 5.2. There could be several explanations for this. One likely explanation is that noise added to the models by DP does not change the characteristics of the model (e.g., the structure of the gradient remains the same). Additionally, the results are heavily dependent on the implementation. While we believe that SmartNoise offers a well-tested implementation, implementation details could affect the quality of the results with respect to our domain-specific metrics. Another noticeable detail is that the individual with ID 4 has the largest dispersion on TP regardless of the choice of ϵ. The reason for this could be that this individual did not provide data for 24 h, which entails data with more measurement gaps than the other participants in the dataset. Furthermore, ID 4 has more tachycardia minutes compared to the other participants presented here. Thus, it could be concluded that the synthetic data generators perform worse on outlier data.

When looking at the privacy values of PATE-CTGAN, there seems to be a downward trend for TP with increasing ϵ. In addition, the dispersion is also smaller for PATE-CTGAN compared to the other methods (see Figure 3). Looking at the mean TP values for small ϵ, it can be seen that PATE-CTGAN has the highest values for $\epsilon = 0.5$ compared with the other techniques. However, overall DP-CTGAN has a better TP for $\epsilon = 1$ and scores the highest TP values there compared to the other techniques. While one might conclude that DP-CTGAN performs best in terms of our TP score, it should be noted that DP-CTGAN has a wider dispersion and in some cases PATE-CTGAN achieves better privacy, particularly with $\epsilon = 0.5$. Additionally, it should be noted that MWEM shows relatively stable TP values and seems to be independent from ϵ. While this is not necessarily the preferred behavior, MWEM is also out of competition because of the reasons mentioned in the beginning of this section.

In terms of the utility metric TU, DP-CTGAN performs better than the other methods up to an epsilon of 5 and can compete thereafter. For smaller values, it can be also seen that PATE-CTGAN cannot keep up and only starts to produce usable results for an ϵ of around 6. This changes for larger ϵ values, which also seems to be a literature consensus that PATE-CTGAN performs better for larger ϵ [11]. The explanation for this phenomenon could be how DP is incorporated in the respective mechanism. While for DP-CTGAN noise is only applied to the weights between training epochs, PATE-CTGAN applies noise to the weights of every teacher network. This could lead to weaker error propagation as ϵ increases.

Furthermore, the privacy–utility trade-off is quantified with our ratio, which indicates that there are clear maxima for the choice of ϵ and that there is a turning point when the increase in utility does not outweigh the loss of privacy anymore. We found that this happens for relatively large ϵ values of approximately 6. While this seems high, there are real-world applications like the US census where even higher ϵ values are considered adequate for privacy protection [37].

Regarding the runtime measurement, it can be observed that the runtime increases with a growing ϵ. Due to the higher ϵ value, the models of the synthetic data generators become more complex, leading to this observation. It should be noted that MWEM is again a special case. The runtime of MWEM is practically independent of ϵ, since no model needs to be trained, but it is based on histograms. It also runs on the CPU instead of the GPU, which makes a direct comparison difficult.

Finally, it can be concluded for $\epsilon < 1$ that PATE-CTGAN performs best in terms of the privacy metric TP. For larger values, neither DP-CTGAN nor PATE-CTGAN can achieve a clear advantage considering the higher dispersion of DP-CTGAN. As mentioned earlier, DP-CTGAN performs best predominantly in terms of the utility Tachycardia Utility (TU), only for very large ϵ values PATE-CTGAN which are slightly better when looking at the range of values between different runs. However, this comes at the cost of performance. Given the results of the privacy utility trade-off, higher ϵ can be considered to improve the usability of the data while still maintaining a decent privacy level. For small ϵ values, DP-CTGAN has a considerable advantage. In general, MWEM does not seem to be competitive for our use case and is therefore not recommended.

7.2. Limitations

It should be emphasized that this work has certain limitations. One main limitation is the focus on the Fitbit dataset containing the heart rate measurements and the data-related use case. While we provide an in-depth look into this use case, there is no generalization of the analyses that address other use cases or more general experiments. Additionally, the choice of attacker in this works is strongly tied to the use case. A more general attacker could cover broader privacy issues. Another limitation is the choice of technologies for this article. Whilst the SmartNoise framework is widely used and very popular, this choice limits the results. Experiments with more and different data generators could be performed in the future. The experiments in this work also rely on the default parameters and settings of SmartNoise. Besides those limitations, this article provides a use case-specific evaluation of the technologies and underlines the need for such an analysis.

8. Conclusions and Future Work

According to our experiments, the use of private synthetic data generators seems very promising. However, the evaluation shows that the original DP parameter ϵ is not sufficient as the only privacy metric and the meaning for privacy and utility still remains an open question. This is especially true for use cases in which time series data are released. Certain features are still preserved by the generative models which can pose a privacy threat. In our example, this can be observed in the heart frequency curve, where a change in the heart rate throughout the day or certain patterns could potentially be linked to an individual's activities. We therefore argue that a use case-specific analysis is required when employing such technologies. For our real-world scenario, we did this by introducing the Tachycardia Utility (TU) and Tachycardia Privacy (TP) metrics, which measure utility and privacy while also considering the specifics of the data. This is illustrated by the fact that these metrics do not necessarily correlate with DP's typical ϵ metric for the technologies used. A variety of experiments using these metrics were conducted with the three synthetic data generators MWEM, PATE-CTGAN, and DP-CTGAN. These experiments include an evaluation of the privacy and utility metrics TP and TU compared to the original notion of ϵ. DP-CTGAN seems to perform best, but there is no clear winner for our benchmarks. Due to its different approach, MWEM is not recommended for scenarios to those in this article. In general, synthetic data generators provide a useful way to generate more accurate and private data than a static data release which uses traditional PETs. Finally, it should be considered that the meaning ϵ for those technologies is not necessarily the only indicator of privacy and that more use-case-specific metrics are recommended.

It remains to be stated that more work in this area of research is needed. Further experiments could include a comparison with interactive DP technologies provided by frameworks such as Google DP (https://github.com/google/differential-privacy (accessed on 5 September 2022)). Additionally, a broader and more diverse set of data generator technologies should be evaluated for the use case presented here. While this future work focuses on the specific use case, the research question should also be posed more generally. It should be tested how robust the metrics introduced herein are for other time series data or if more general definitions are needed. Furthermore, the privacy attacker presented herein is very specific and should be defined more universally. Another question is that of how models could look that seek to preserve certain features through constraints defined by medical background knowledge while simultaneously providing privacy protection. This includes the question of how strictly consistency, structural, and correlation constraints could be defined for medical data. Such a technology would be even more useful if it could work with categorical values so that it could, for example, create synthetic patients' data with various disease types. Ultimately, this could lead to the vision of a private synthetic patient generator which generates accurate but private data based on real-world patients.

Author Contributions: Conceptualization, A.A. and M.L.; methodology, A.A. and M.L.; software, M.L.; investigation, M.L. and A.A.; writing—original draft preparation, A.A. and M.L.; writing—review and editing, A.A., P.P., M.L. and E.K.; supervision, P.P., E.K. and J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported as a Fraunhofer Lighthouse Project and was conducted in close cooperation with the KASTEL project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: The code for the experiments and evaluation of this work is provided here: https://gitlab.cc-asp.fraunhofer.de/synthetic-data-generation/medical-data-analyses (accessed on 26 November 2022).

Abbreviations

The following abbreviations are used in this manuscript:

DP	Differential Privacy
ePA	Elektronische Patientenakte
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
MWEM	Multiplicative Weights Exponential Mechanism
PETs	Privacy-Enhancing Technologies
PIF	Privacy Impact Factor
TP	Tachycardia Privacy
TU	Tachycardia Utility

References

1. Appenzeller, A. Privacy and Patient Involvement in e-Health Worldwide: An International Analysis. In *Proceedings of the 2020 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory*; Beyerer, J., Zander, T., Eds.; KIT Scientific Publishing: Karlsruhe, Germany, 2021; Volume 51, pp. 1–17. [CrossRef]
2. Martin-Sanchez, F.J.; Aguiar-Pulido, V.; Lopez-Campos, G.H.; Peek, N.; Sacchi, L. Secondary Use and Analysis of Big Data Collected for Patient Care: Contribution from the IMIA Working Group on Data Mining and Big Data Analytics. *Yearb. Med. Informatics* **2017**, *26*, 28–37. [CrossRef] [PubMed]
3. Sweeney, L. k-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness-Knowl. Based Syst.* **2002**, *10*, 557–570. [CrossRef]
4. Fung, B.C.M.; Wang, K.; Chen, R.; Yu, P.S. Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Comput. Surv.* **2010**, *42*, 1–53. [CrossRef]
5. Alvim, M.S.; Andrés, M.E.; Chatzikokolakis, K.; Degano, P.; Palamidessi, C. Differential Privacy: On the trade-off between Utility and Information Leakage. In *International Workshop on Formal Aspects in Security and Trust*; Springer: Berlin/Heidelberg, Germany, 2011. [CrossRef]
6. Abay, N.C.; Zhou, Y.; Kantarcioglu, M.; Thuraisingham, B.; Sweeney, L. Privacy Preserving Synthetic Data Release Using Deep Learning. In Proceedings of the Machine Learning and Knowledge Discovery in Databases, Dublin, Ireland, 10–14 September 2018; Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 510–526. [CrossRef]
7. Stadler, T.; Oprisanu, B.; Troncoso, C. Synthetic Data—Anonymisation Groundhog Day. 2020. Available online: https://arxiv.org/abs/2011.07018 (accessed on 5 July 2022).
8. Bellovin, S.M.; Dutta, P.K.; Reitinger, N. Privacy and Synthetic Datasets. *SSRN Electron. J.* **2018**, *22*, 1. [CrossRef]
9. Chen, R.J.; Lu, M.Y.; Chen, T.Y.; Williamson, D.F.K.; Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* **2021**, *5*, 493–497. [CrossRef] [PubMed]
10. Bowen, C.M.; Snoke, J. Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge. 2019. Available online: https://arxiv.org/abs/1911.12704 (accessed on 8 July 2022).
11. Rosenblatt, L.; Liu, X.; Pouyanfar, S.; de Leon, E.; Desai, A.; Allen, J. Differentially Private Synthetic Data: Applied Evaluations and Enhancements. 2020. Available online: https://arxiv.org/abs//2011.05537 (accessed on 8 July 2022).
12. Ping, H.; Stoyanovich, J.; Howe, B. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, 27–29 June 2017; Association for Computing Machinery: New York, NY, USA, 2017. [CrossRef]
13. Al Aziz, M.M.; Ahmed, T.; Faequa, T.; Jiang, X.; Yao, Y.; Mohammed, N. Differentially Private Medical Texts Generation Using Generative Neural Networks. *ACM Trans. Comput. Healthc.* **2021**, *3*, 1–27. [CrossRef]
14. Carlini, N.; Liu, C.; Erlingsson, Ú.; Kos, J.; Song, D. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In Proceedings of the 28th USENIX Security Symposium (USENIX Security 19), Santa Clara, CA, USA, 14–16 August 2019; USENIX Association: Berkeley, CA, USA, 2019; pp. 267–284. Available online: https://www.usenix.org/conference/usenixsecurity19/presentation/carlini (accessed on 26 November 2022).
15. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. ℓ-diversity: Privacy Beyond k-Anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3. [CrossRef]

16. Samarati, P.; Sweeney, L. *Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression*; Technical Report SRI-CSL-98-04; Computer Science Laboratory, SRI International: Menlo Park, CA, USA, 1998. Available online: http://www.csl.sri.com/papers/sritr-98-04/ (accessed on 26 November 2022).
17. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 17–20 April 2007; pp. 106–115. [CrossRef]
18. Nergiz, M.E.; Atzori, M.; Clifton, C. Hiding the presence of individuals from shared databases. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data—SIGMOD '07, Beijing, China, 11–14 June 2007; ACM Press: Beijing, China, 2007; pp. 665–676. [CrossRef]
19. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 265–284. [CrossRef]
20. Dwork, C.; Roth, A. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* **2013**, *9*, 211–407. [CrossRef]
21. Rubin, D.B. Statistical disclosure limitation. *J. Off. Stat.* **1993**, *9*, 461–468.
22. Li, S.C.; Tai, B.C.; Huang, Y. Evaluating Variational Autoencoder as a Private Data Release Mechanism for Tabular Data. In Proceedings of the 2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC), Kyoto, Japan, 1–3 December 2019; pp. 198–206. [CrossRef]
23. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. 2014. Available online: https://arxiv.org/abs/1406.2661 (accessed on 24 June 2022).
24. Kopp, A. Microsoft SmartNoise Differential Privacy Machine Learning Case Studies. 2021. Available online: https://azure.microsoft.com/en-us/resources/microsoft-smartnoisedifferential-privacy-machine-learning-case-studies/ (accessed on 14 April 2022).
25. Hardt, M.; Ligett, K.; McSherry, F. A Simple and Practical Algorithm for Differentially Private Sata Release. 2010. Available online: https://arxiv.org/abs/1012.4763 (accessed on 1 July 2022).
26. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular Data Using Conditional GAN. 2019. Available online: https://arxiv.org/abs/1907.00503 (accessed on 29 June 2022).
27. Papernot, N.; Song, S.; Mironov, I.; Raghunathan, A.; Talwar, K.; Erlingsson, U. Scalable Private Learning with PATE. 2018. Available online: https://arxiv.org/abs/1802.08908 (accessed on 29 June 2022).
28. Centers for Medicare & Medicaid Services. *The Health Insurance Portability and Accountability Act of 1996 (HIPAA)*; Centers for Medicare & Medicaid Services: Baltimore, MD, USA, 1996.
29. Janmey, V.; Elkin, P.L. Re-Identification Risk in HIPAA De-Identified Datasets: The MVA Attack. In Proceedings of the AMIA 2018, American Medical Informatics Association Annual Symposium, San Francisco, CA, USA, 3–7 November 2018. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371259/ (accessed on 26 November 2022).
30. de Montjoye, Y.A.; Hidalgo, C.A.; Verleysen, M.; Blondel, V.D. Unique in the Crowd: The privacy bounds of human mobility. *Sci. Rep.* **2013**, *3*, 1376. [CrossRef] [PubMed]
31. Narayanan, A.; Shmatikov, V. How to Break Anonymity of the Netflix Prize Dataset. 2006. Available online: https://arxiv.org/abs/cs/0610105 (accessed on 7 June 2022).
32. Deußer, C.; Passmann, S.; Strufe, T. Browsing Unicity: On the Limits of Anonymizing Web Tracking Data. In Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP), Virtual, 18–20 May 2020; pp. 777–790. [CrossRef]
33. WHO. ICD10 Code for Tachycardia. 2019. Available online: https://icd.who.int/browse10/2019/en#/R00.0 (accessed on 31 March 2022).
34. Furberg, R.; Brinton, J.; Keating, M.; Ortiz, A. Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016 [Data set]. 2016.
35. Dankar, F.K.; El Emam, K. Practicing Differential Privacy in Health Care: A Review. *Trans. Data Priv.* **2013**, *6*, 35–67. Available online: https://www.tdp.cat/issues11/tdp.a129a13.pdf (accessed on 26 November 2022).
36. Zhang, S.; Hagermalm, A.; Slavnic, S. An Evaluation of Open-Source Tools for the Provision of Differential Privacy. 2022. Available online: https://arxiv.org/abs/2202.09587 (accessed on 6 July 2022).
37. Garfinkel, S. Differential Privacy and the 2020 US Census. In *MIT Case Studies in Social and Ethical Responsibilities of Computing*; Winter; MIT: Cambridge, MA, USA, 2022. [CrossRef]