Accepted Manuscript online: 2024-08-13 Article published online: 2024-09-09

Does Differentially Private Synthetic Data Lead to Synthetic Discoveries?

Ileana Montoya Perez[1] Parisa Movahedi[1] Valtteri Nieminen[1] Antti Airola[1] Tapio Pahikkala[1]

[1] Department of Computing, University of Turku, Turku, Finland

Methods Inf Med

Address for correspondence Ileana Montoya Perez, PhD, Department of Computing, University of Turku, 452B, Agora 4th floor, Vesilinnantie 5, 20500, Turku, Finland (e-mail: iimope@utu.fi).

Abstract

Background Synthetic data have been proposed as a solution for sharing anonymized versions of sensitive biomedical datasets. Ideally, synthetic data should preserve the structure and statistical properties of the original data, while protecting the privacy of the individual subjects. Differential Privacy (DP) is currently considered the gold standard approach for balancing this trade-off.

Objectives The aim of this study is to investigate how trustworthy are group differences discovered by independent sample tests from DP-synthetic data. The evaluation is carried out in terms of the tests' Type I and Type II errors. With the former, we can quantify the tests' validity, i.e., whether the probability of false discoveries is indeed below the significance level, and the latter indicates the tests' power in making real discoveries.

Methods We evaluate the Mann–Whitney U test, Student's *t*-test, chi-squared test, and median test on DP-synthetic data. The private synthetic datasets are generated from real-world data, including a prostate cancer dataset ($n = 500$) and a cardiovascular dataset ($n = 70{,}000$), as well as on bivariate and multivariate simulated data. Five different DP-synthetic data generation methods are evaluated, including two basic DP histogram release methods and MWEM, Private-PGM, and DP GAN algorithms.

Conclusion A large portion of the evaluation results expressed dramatically inflated Type I errors, especially at levels of $\varepsilon \leq 1$. This result calls for caution when releasing and analyzing DP-synthetic data: low *p*-values may be obtained in statistical tests simply as a byproduct of the noise added to protect privacy. A DP Smoothed Histogram-based synthetic data generation method was shown to produce valid Type I error for all privacy levels tested but required a large original dataset size and a modest privacy budget ($\varepsilon \geq 5$) in order to have reasonable Type II error levels.

Keywords

- differential privacy
- synthetic data
- hypothesis testing
- statistical inference
- Mann–Whitney U test

Introduction

As the amount of health and medical data collected from individuals has grown, so has the interest in using it for secondary purposes such as research and innovation. Many benefits have been proposed to arise from sharing these data,[1] for example, enhancing research reproducibility, building on existing research, performing meta-analyses, and reducing clinical trial costs by reusing existing data. However, privacy concerns about the potential harm to individuals that may come from sharing their sensitive

received
May 31, 2023
accepted after revision
July 25, 2024
accepted manuscript online
August 13, 2024

DOI https://doi.org/10.1055/a-2385-1355.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

data, along with legislation aimed at addressing these concerns, restrict the opportunities for sharing individuals' data.

The release of synthetic data, generated using a statistical model derived from an original sensitive dataset, has been proposed as a potential solution for sharing biomedical data while preserving individuals' privacy.[2–4] It has been argued that since synthetic data consist of synthetic records instead of actual records, and synthetic records are not associated with any specific identity, privacy is preserved.[2] However, it has been repeatedly demonstrated that this is not the case as synthetic data are not inherently privacy-preserving.[5–8] In the worst case, a generative model could create near copies of the original sensitive data it was trained on. Moreover, there are many more subtle ways that models can leak information about their training data.[5,9] At the other extreme, perfect anonymity is guaranteed only when no useful information from the original data remains. Therefore, in addition to preserving privacy, the generated data should have high utility, meaning the degree to which the inferences obtained from the synthetic data correspond to inferences obtained from the original data.[5,10] Consequently, when generating synthetic data, it is essential to find a balance between the privacy and utility of the data, ensuring that the generated data capture the primary statistical trends in the original data while also preventing the disclosure of sensitive information about individuals.[11]

Differential Privacy (DP), a mathematical formulation that provides probabilistic guarantees on the privacy risk associated with disclosing the output of a computational task, has been widely accepted as the gold standard of privacy protection.[12–15] As a result, methods that ensure DP guarantees have been introduced in a broad range of settings, including descriptive statistics,[13,16] inferential statistic,[17–20] and machine learning applications.[15,21] Furthermore, DP offers a theoretically well-founded approach that provides probabilistic privacy guarantees also for the release of synthetic data. Therefore, several methods for releasing DP-synthetic data have been proposed.[22–26] Some state-of-the-art methods for generating DP-synthetic data use multi-dimensional histograms, which are standard tools for estimating the distribution of data with minimal a priori assumptions about its statistical properties. Other methods are based on machine learning–based generative models, for example, Bayesian and Generative Adversarial Network (GAN)-based methods. The aim of DP-synthetic data is to be a privacy-preserving version of the original data that could be safely used in its place, requiring no expertise on DP or changes to the workflow from the end-user. However, DP-synthetic data are always a distorted version of the original data, and especially when high levels of privacy are enforced the level of distortion can be quite considerable. Even though combining DP with synthetic data guarantees a desired level of privacy, preservation of the utility remains unclear. In particular, the validity of statistical significance tests, namely the statistical guarantees of the false-finding probabilities being at most the significance level, may be lost.

Hypothesis tests for assessing whether two distributions share a certain property are essential tools in analyzing biomedical data. In this work, we particularly focus on the Mann–Whitney (MW) U test (a.k.a. Wilcoxon rank-sum test or Mann–Whitney–Wilcoxon test), as it is the de facto standard for testing whether two groups are drawn from the same distribution.[27,28] It is widely applied in medical research,[29] particularly when analyzing a biomarker between nonhealthy and healthy patients in clinical trials. It is well known that the MW U test is valid for this question, that is, the probability of falsely rejecting the null hypothesis of the two groups being drawn from the same distribution is at most the significance level determined a priori.[30] Alongside the MW U test, we also consider the Student's *t*-test,[31] median test,[32] and chi-squared test.[33] In general, the choice of statistical test should be guided by the distribution characteristics of the dataset and the datatype under analysis.

In order for DP-synthetic data to be useful for basic use cases in medical research, such as the MW U test, one would hope to observe roughly similar results when carrying out tests on sensitive medical datasets. Otherwise, there is a risk that discoveries are missed because of information lost in synthetization, or worse, that false discoveries are made due to artifacts introduced in the data generation process.

Objectives

DP-synthetic data have been proposed as a solution for publicly releasing anonymized versions of sensitive data such as medical records. Ideally, this would allow for performing reliable statistical analyses on the DP-synthetic data without ever needing to access the original data (see ►**Fig. 1**). However, there is a risk that DP-synthetic data generation methods distort the original data in ways that can lead to loss of information and even to false discoveries.

In this study, we empirically evaluate the validity and power of independent sample tests, such as the MW U test, applied to DP-synthetic data. The Type I and Type II errors are used to measure the test validity and power, respectively. On one hand, a test is valid if, for any significance level, the probability that it falsely rejects a correct null hypothesis is no larger than the significance level.[34] If the test is not valid, its use can lead to false scientific discoveries, and hence its practical utility can be even negative. On the other hand, the test's power refers to the probability of correctly rejecting a false null hypothesis.

In our experiments with the MW U test, we evaluated five different DP-synthetic data generation methods on bivariate real-world medical datasets, as well as data drawn from two Gaussian distributions. Additionally, we performed experiments with simulated multivariate data to explore the behavior of MW U test, Student's *t*-test, median test, and chi-squared test on higher dimensional DP-synthetic data consisting of different variable types. Our study contributes to understanding the reliability of statistical analysis when DP-synthetic data are used as a proxy for private data whose public release is challenging or even impossible.

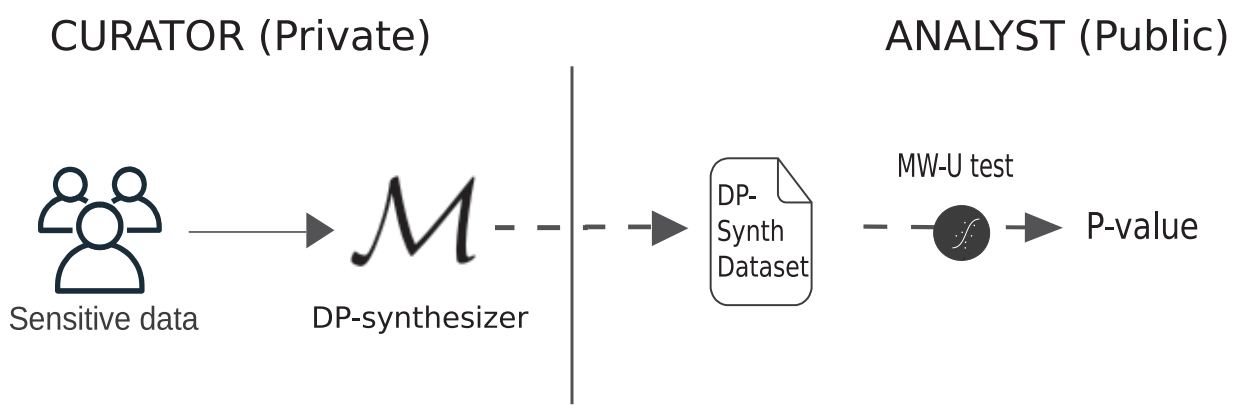

Fig. 1 The overall configuration of the study.

Methods

In this section, we first present the formal definition of DP. Next, we introduce DP methods for synthetic data generation while describing the five DP-synthesizers used in this study. Following that, we explain the validity and power of a statistical test. Finally, we introduce the independent sample tests considered in this study.

Differential Privacy

DP is a mathematical definition that makes it possible to quantify privacy.[12,13] A randomized algorithm M satisfies (ε, δ)-DP if for all outputs S of M and for all possible neighboring datasets D, D' that differ by only one row,

$$Pr[M(D) = S] \leq e^{\epsilon} Pr[M(D') = S] + \delta \quad (1)$$

where ε is an upper bound on the privacy loss, and δ is a small constant corresponding to a small probability of breaking the DP constraints. For $\delta = 0$ in particular, solving (1) w.r.t. ε results to:

$$\log(Pr[M(D) = S]) - \log(Pr[M(D') = S]) \leq \epsilon \quad (2)$$

indicating that the log-probability of any output can change no more than ε. Accordingly, an algorithm M which is ε-DP guarantees that for every run $M(D)$, the outcome obtained is almost equally likely to be obtained on any neighboring dataset, bounded by the value of ε. Informally, in DP, privacy is understood to be protected at a given level of ε if the algorithm's output does not overly depend on the input data of any single contributor; it should yield a similar result if the individual's information is present in the input or not.

Typically, DP methods are nonprivate methods that are transformed to fulfill the DP definition. This is achieved by adding noise using a noise mechanism calibrated based on the ε and the algorithm to be privatized.[12,13] Choosing the appropriate value of epsilon is context-specific and an open question, but, for example, $\varepsilon \leq 1$ has been proposed to provide a strong guarantee,[35] while $1 < \varepsilon \leq 10$ is considered to still produce useful privacy guarantees,[36] depending on the task and type of data.

DP Methods for Synthetic Data Generation

In recent years, several methods for generating DP-synthetic data have been proposed.[23,24,26,37,38] Some of the proposed methods are based on histograms or marginals. These methods privatize the cell counts or proportions of a cross-tabulation of the original sensitive data to generate the DP-synthetic data. Other methods use a parameterized distribution or a generative model that has been privately derived from the original data. While DP methods based on histograms or marginals have been found to produce usable DP-synthetic data with a reasonable level of privacy guarantee, methods based on parameterized distributions or deep learning–based generative models have presented greater challenges.[39,40]

Generative methods based on marginals share a three-step process: initially, a set of marginals is identified, either manually by a domain expert or through DP-automatic selection. Next, these chosen marginals are measured using DP. Finally, synthetic data are generated from the noisy marginals. To address the challenges of high-dimensional domains, recent methods have been developed to automatically and privately select a subset of marginals ensuring their preservation in the synthetic data generated, such as PrivMRF,[41] PrivBayes,[42] MWEM (Multiplicative Weights Exponential Mechanism),[22] and AIM.[43]

PrivMRF employs Markov Random Fields to generate synthetic data under DP, emphasizing the retention of statistical correlations between selected marginals within the privacy constraints. PrivBayes constructs a Bayesian network under DP, utilizing a selected set of marginals to approximate the underlying data distribution for synthetic data generation. The MWEM algorithm is designed to generate a data distribution that yields query responses closely resembling those obtained from the actual dataset. AIM on the other hand is a workload-adaptive algorithm, allowing for the input of a predefined set of marginals to be specifically preserved in the final DP-approximated distribution.

There are two approaches to consider when designing a DP workflow: global DP and local DP.[13] Global DP involves aggregating data in a central location and is managed by a trusted curator, ensuring privacy at the dataset level. In contrast, local DP decentralizes the privacy mechanism by applying it directly to the individual's data before it is shared.

Many applications, such as crowdsourced systems, involve data distributed across multiple individuals who do not trust any other party. These individuals are only willing to share their information if it has been privatized on their own devices prior to transmission. In such cases, local privacy methods such as LoPub and LoCop become applicable, ensuring that each individual's data remain confidential even when aggregated from diverse sources.[44–46]

In this study, we focus on five well-known DP methods for generating synthetic data in a global DP setting. These methods have established algorithms or available packages, making them accessible to any practitioner. Following, we provide a brief description of each of these DP methods.

- **DP Perturbed Histogram**
  This method uses the Laplace mechanism[13] to privatize the original histogram bin counts. The noise added to each bin is sampled separately from a calibrated Laplace distribution. After adding the noise, all negative counts are set to zero, and individual-level data are generated from the noisy counts.
- **DP Smoothed Histogram**
  This method generates synthetic data by randomly sampling from the probability distribution determined by the following histogram. The probabilities of the histogram bins are proportional to $c_i + 2m/\varepsilon$, where c_i is the number of original data points in the ith histogram bin and m is the size of the synthetic data drawn. The approach is similar to the one discussed by Wasserman and Zhou.[14] Unlike the other considered DP methods, the utility of this method is inversely proportional also to the amount synthetic data drawn. Therefore, in our experiments, we use the method only in settings where the size of the synthetic data generated is considerably smaller than that of the original sensitive data. A proof of the approach being DP is presented in ►Supplementary Material A.1.
- **Multiplicative Weights Exponential Mechanism**
  This algorithm proposed by Hardt et al[22] is based on a combination of the multiplicative weights update rule with the exponential mechanism. The MWEM algorithm estimates the original data distribution using a DP iterative process. Here, a uniform distribution over the variables of the original data is updated using the multiplicative weighting of a query or bin count selected through the exponential mechanism and privatized with the Laplace mechanism in each iteration. The privacy budget ε is split by the number of iterations, as in every iteration the original data need to be accessed.
- **Private-PGM**
  McKenna et al[26] propose this approach for DP-synthetic data generation. It consists of three basic steps: (1) selecting a set of low-dimensional marginals (i.e., queries) from the original data. (2) Adding calibrated noise to the marginals. (3) Generating synthetic data that best explain the noisy marginals. In step 3, based on the noisy marginals, a Probabilistic Graphical Model (PGM) is used to determine the data distribution that best captures the variables' relationship and enables synthetic data generation.
- **Differentially Private GAN**
  GANs[47] consist of a generator, denoted with **G**, and one or more discriminators **D**. The goal is that **G** would learn to produce synthetic data similar to the original data. The two networks are initialized randomly and trained iteratively in a game-like setup: **G** is fed noise to create synthetic data, which the **D** tries to discriminate as being original or synthetic. The generator uses feedback from the discriminator(s) to update its parameters via gradient descent (see Goodfellow[48] for a detailed explanation). GANs, and other deep learning models, can attain privacy guarantees by using a DP version of an optimization algorithm, most often differentially private stochastic gradient descent.[36]

Validity and Power of Independent Sample Tests

Samples are considered independent when individuals in one group do not influence or share information with individuals in another group. Each group consists of unique members, and no pairing or matching occurs between them. To evaluate potential statistical differences between the two groups, researchers commonly use statistical tests designed for independent samples. These tests determine whether the samples were drawn independently from distributions with shared properties. The independent sample tests considered in this work are the MW U test, Student's t-test, median test, and chi-squared test.

The validity and power of a statistical test can be evaluated in terms of Type I and Type II errors. Let us recall that Type I is the error incurred when a "True" null hypothesis is rejected, producing false inference. On the other hand, Type II is the error of failing to reject a "False" null hypothesis (see ►Fig. 2). Following Casella and Berger,[34] we say that p-value corresponding to the observed test statistic is valid if it is at most the probability of observing as extreme test statistic under the null hypothesis. Consequently, the significance test is valid if its p-value is valid.

A priori selected significance level α defines a threshold that, for any valid hypothesis test, forms an upper bound on the probability of committing Type I error. A typical choice for α is 0.05, indicating a maximum 5% chance of incorrectly rejecting a true null hypothesis. The probability of making a Type II error is often denoted as β (beta), from which the power of the test can be determined by computing $1 - \beta$. The power of a test can be interpreted as the probability of correctly rejecting a null hypothesis when it is in fact "False." The power depends on the analysis task, being affected by factors such as chosen significance level, the effect size, the sample size, and the relative sizes of the different groups. In our experiments, we observed the imbalance between group sizes to have a dominant effect on tests' power in practice, because of the DP-synthetic data generators' tendency to produce imbalanced samples for small ε values.

Mann–Whitney U test

The MW U test is a statistical test first proposed by Frank Wilcoxon in 1945 and later, in 1947, formalized by Henry Mann and Donald Whitney.[49,50] While there are many

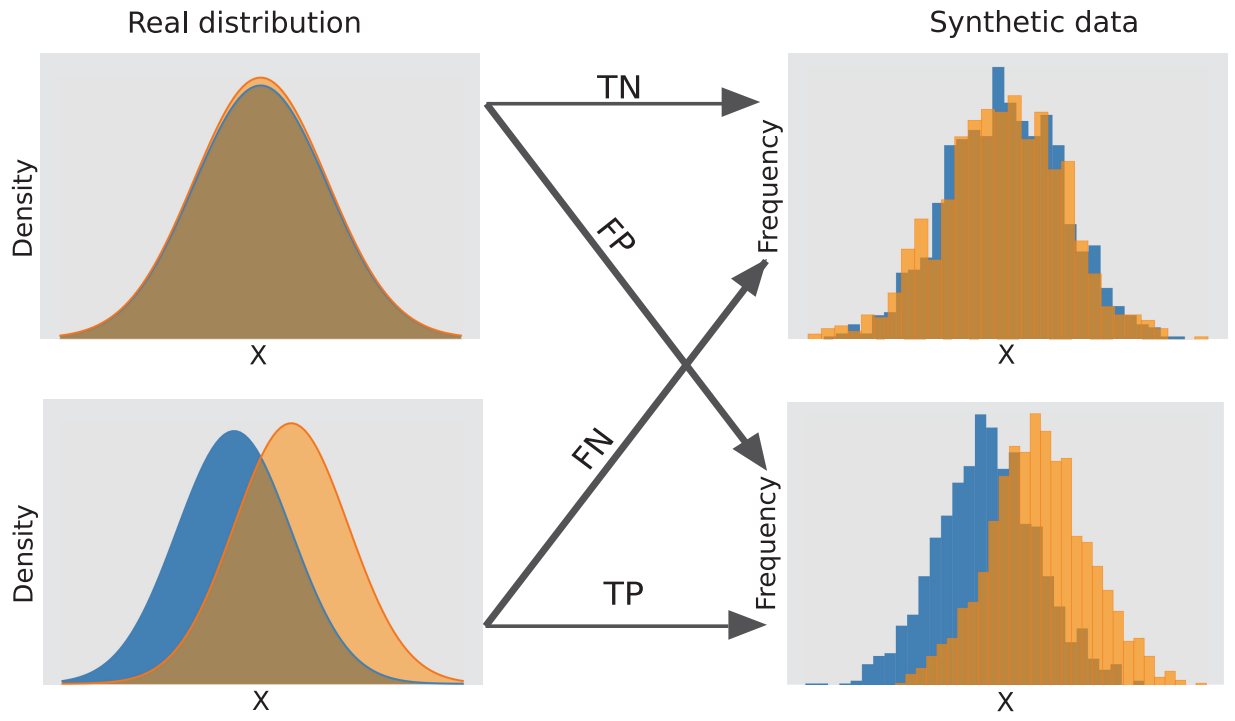

Fig. 2 Possible outcomes of a hypothesis test that tests whether two distributions are the same. FN, false negative (Type II error); FP, false positive (Type I error); TN, true negative; TP, true positive.

different uses and interpretations of the test (see, e.g., Fay and Proschan[30] for a comprehensive review), in this article we focus on the null hypothesis that two samples or groups are drawn from the same distribution. The test carried out on two groups produces a value of the MW U statistic and the corresponding p-value. The U statistic measures the difference between the groups as the number of times an observed member of the first group is smaller than that of the second group, ties being counted as a half-time. The p-value indicates the strength of evidence the value of the U-statistic provides against the null hypothesis, given that the assumption of the data being independently drawn holds.[34]

Couch et al[19] proposed a differentially private version of the MW U test (DP-MW). The DP-MW U test is presented as (ε, δ)-DP, where a portion of the privacy budget ε and δ are used for privatizing the smallest group size. The privatized size and the rest of ε are then used for privatizing the U statistic using a calibrated Laplace distribution. In order to calculate the corresponding p-value, the DP-MW U distribution under the null hypothesis is generated based on the privatized group sizes. Detailed information and algorithms are provided by Couch et al.[19] The DP-MW U test is not based on analyzing synthetic data, but rather the test is carried out directly on the original sensitive dataset, and DP guarantees that sensitive information about individuals is not leaked when releasing the test results.

In this study, the DP-MW U test on the original sensitive data provides us with a reference point, a valid test with the best-known achievable power when performing MW U test under DP. In contrast, the ordinary MW U test is evaluated on the DP-synthetic data. If the validity of the ordinary test is preserved, comparison to the reference point indicates how much power is lost when general-purpose DP-synthetic data are generated as an intermediate step.

Student's *t*-Test, Chi-Squared Test, and Median Test

The Student's t-test (independent or unpaired t-test) is a widely utilized parametric statistical test that assesses whether the means of two independent samples are significantly different.[31] The null hypothesis states that the means are statistically equivalent, while the alternative hypothesis suggests that they are not. The test is valid for two independent samples if their distributions are normal and their variances are equal.

The chi-squared test is a nonparametric test used to analyze the association of two categorical variables by utilizing a contingency table.[33] Under the null hypothesis, the observed (joint) frequencies should equal to expected (marginal) frequencies, meaning that the variables are independent. Since, under the null hypothesis, the test statistic approximately follows a chi-squared distribution, the test's validity depends on the sample size. However, for small n and for 2×2 tables, the appropriate alternative is the Fisher's exact test.

Median test is a nonparametric method used to test the null hypothesis of two (or more) independent samples being drawn from distributions of equal medians.[32] The test is valid as long as the distributions have equal densities in the neighborhood of the median (see, e.g., Freidlin and Gastwirth[51] and reference therein).

Experimental Evaluation

To empirically evaluate the utility of independent sample tests applied to DP-synthetic datasets, we conducted a set of

experiments. In each experiment, either simulated or real-world data were used to represent the original sensitive dataset. These data were subsequently used to train DP-synthetic data generation methods. Finally, the independent sample tests were carried out on synthetic data produced by the generator.

First, we examined the behavior of MW U test on DP-synthetic data generated based on bivariate real-world datasets or simulated data drawn from Gaussian distributions. As depicted by real distribution of ►Fig. 2, we considered two cases for Gaussian data: one where both groups are drawn from the same distribution (i.e., the null hypothesis is true) and one where they are drawn from distributions with different means (i.e., the null hypothesis is false). While in practice synthesizing datasets consisting of only two variables would have quite limited use cases, these experiments allow demonstrating the fundamentals of how different DP synthetization approaches affect the validity and power of statistical tests. In order to provide a more realistic setup, we further performed experiments on a simulated multivariate dataset. The validity and power of the MW U test, Student's *t*-test, median test, and chi-squared test were explored in these experiments.

In the overall study design (see ►Fig. 1), the real-world, Gaussian, and simulated multivariate datasets correspond to the sensitive data given as input to a DP-synthesizer method that produces a DP-synthetic dataset as output. In the following subsections, we present the datasets, the implementation details of the DP-synthetic data generation methods used, and the experiments conducted.

Original Datasets

First, we experimented with a setup, where the sensitive original data consist of only two variables (i.e., a binary variable and a continuous variable). The binary variable represents group membership (e.g., healthy or nonhealthy), while the continuous variable is the one used to compare the groups with the MW U test.

To establish a controlled environment where the amount of signal (i.e., the effect size) in the population is known, we drew two groups of data from two Gaussian distributions with a known mean (μ) and standard deviation (σ). More precisely, for non-signal data, which corresponds to a setting where the null hypothesis is true, the two groups were randomly drawn from the same Gaussian distribution ($\mu = 50$, $\sigma = 2$). For the signal data, which correspond to a setting where the null hypothesis is false, two Gaussian distributions with effect size $\mu_1 - \mu_2 = \sigma$ (i.e., $\mu_1 = 51$, $\sigma_1 = 1$, $\mu_2 = 50$, $\sigma_2 = 1$) were used to sample each group. Additionally, for those DP methods based on histograms or marginals, the sampled values for each group were discretized into 100 bins (ranging from 1 to 100).

In order to verify our experiment's results on the Gaussian data, we also carried out experiments using real-world medical data. In this case, we use the following two datasets:

- **The Prostate Cancer Dataset**
  The data are from two registered clinical trials, IMPROD[52] and MULTI-IMPROD,[53] with trial numbers NCT01864135 and NCT02241122, respectively. These trials were approved by the Institutional Review Board, and each enrolled patient gave written informed consent. The dataset consists of 500 prostate cancer (PCa) patients (242 high-risk and 258 benign/low-risk PCa) with clinical variables, blood biomarkers, and magnetic resonance imaging features. For our experiments, we selected two variables: a binary label that indicates the condition of the patient and the prostate-specific antigen (PSA) level. The PSA is a continuous variable that has been associated with the presence of PCa.[54,55] Therefore, in this study, we considered the null hypothesis under test to be "*The PSA levels of high-risk and benign/low-risk PCa patients originate from the same distribution.*" ►Fig. 3A presents the PSA distribution for both groups in this dataset. In those DP methods based on histograms or marginals, the PSA values were discretized using a 40-bin histogram (ranging from 1 to 40, where PSAs ≥ 40 are in the last bin).
- **Kaggle Cardiovascular Disease Dataset**
  This dataset is publicly available and consists of 70,000 subjects and 12 variables, where the target variable is the cardio condition of the subjects, with 34,979 presenting cardiovascular disease and 35,021 without the disease.[56] For our experiments, we use each subject body mass index (BMI), calculated from their weight and height, which has been related to cardiovascular conditions.[57] Here, the null hypothesis under test is "*The BMI level for individuals with the presence of cardiovascular disease and the ones with absence cardiovascular disease originate from the same distribution.*" ►Fig. 3B presents the BMI distribution of both groups (i.e., cardio disease vs. no cardio disease). The BMI variable was discretized into 24 bins, where the first bin contains BMI < 18 and the last bin BMI ≥ 40, in those DP methods that require it.

Finally, we experimented with simulated multivariate datasets. The simulation was based on the real-world PCa dataset. The included variables were the patient's age, PSA level, prostate volume, the use of 5-alpha-reductase inhibitor (5-ARI) medication, prostate imaging reporting and data systems (PIRADS) score, and a class label indicating low-risk or high-risk PCa. The simulated datasets were generated by a GaussianCopulaSynthesizer from the Synthetic Data Vault (SDV)[58] trained on the real-world dataset. In the SVD settings, the age variable was configured to follow a normal distribution, while the remaining numerical variables were set to follow a beta distribution. In experiments with a false null hypothesis, SVD was conditioned to generate simulated datasets with an equal number of high-risk and low-risk patients. For experiments with a true null hypothesis, the condition was to generate only one class (low-risk) for the simulated dataset, and subsequently, half of the data were randomly assigned to the high-risk class.

Implementations

In our experiments, for the generated DP-synthetic data, we used the hypothesis tests provided by the Scipy v1.6.3 package,[59] such as the *mannwhitneyu* function for the

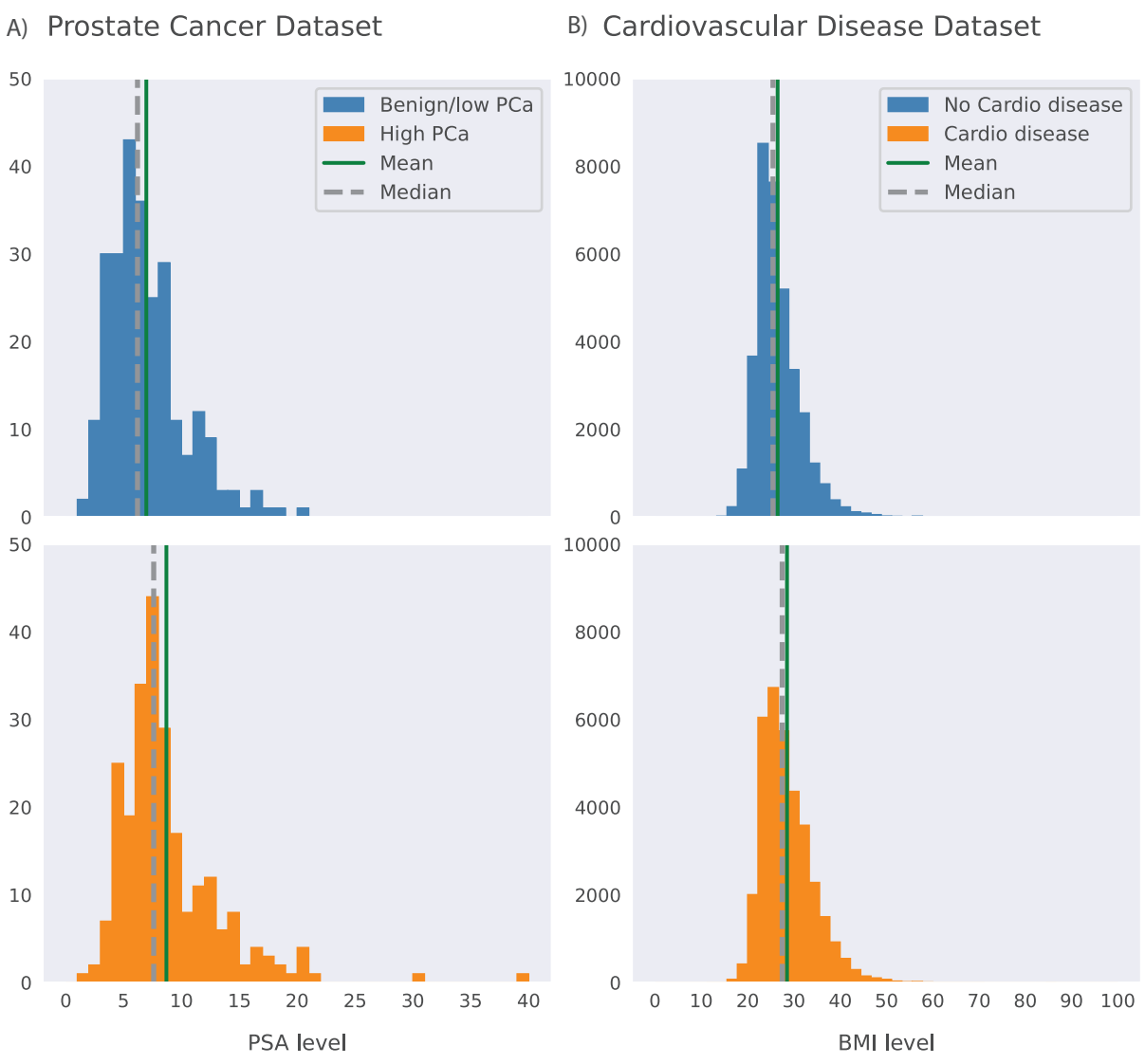

Fig. 3 (**A**) Prostate cancer (PCa) dataset: prostate-specific antigen (PSA) level distribution for high-risk and benign/low PCa. The difference between the groups is statistically significant (MW U stat = 22,713, *p*-value = 1.4e−07), (**B**) Kaggle Cardiovascular Disease dataset: body mass index (BMI) distribution for subjects with the absence and presence of cardio disease. The difference between the groups is statistically significant (MW U stat = 471,500,929.50, *p*-value ≅ 0.000).

MW U test. We used two-sided tests, with all the tests statistics and *p*-values computed using the Scipy function's default values. As a point of reference, we also computed the DP-MW U statistic and *p*-value on the corresponding original sensitive dataset. The DP-MW U test was implemented using Python v3.7 and following the algorithms presented by Couch et al,[19] where 65% of ε and $\delta = 10^{-6}$ are used for estimating the size of the smallest group, and the U statistic is privatized using the estimated size and the remaining ε.

In the case of the DP Perturbed Histogram, Python v3.7 was also used in the implementation. The noise, added to the original histogram, was sampled from a discrete Laplacian distribution[60] scaled by $\frac{2}{\epsilon}$, then the noisy counts were normalized by the original data size. After that, the synthetic data were obtained by transforming the DP histogram counts to values using the bin center point. For Private-PGM[61] and MWEM,[62] their corresponding open-source packages were used to generate DP-synthetic data. The Private-PGM synthetic data were generated by following the demonstration in Python code presented by McKenna et al[26] using Laplace distribution scaled by $\frac{2}{\text{split}(\epsilon)}$ where *split* (ε) is the privacy budget (ε) divided by the number of marginal queries selected. MWEM was run with default hyperparameters; only ε was changed to show the effect of different privacy budgets. The resulting DP-synthetic data were sampled using the histogram noisy weights returned by the MWEM algorithm. The implementation of DP Smoothed Histogram was also coded in Python v.3.7 following Algorithm 1 provided in the ►**Supplementary Material**. In all our experiments, the DP-synthetic data generators were configured to preserve all the one-way and two-way marginals.

The GAN model used is based on the GS-WGAN by Chen et al.[25] The implementation is a modification of the freely available source code,[63] with changes made to suit tabular data

generation instead of images. The generator architecture was changed from a convolutional- to a fully connected three-layer network, and the gradient perturbation procedure was modified to accommodate these changes along with making the source code compatible with an up-to-date version of PyTorch (v1.10.2).[64] Hyperparameter settings were chosen based on the recommendations of Gulrajani et al[65] on the WGAN-GP, which of the GS-WGAN is a DP extension. This model uses privacy amplification by subsampling,[25] a strategy to achieve stronger privacy guarantees by splitting training data into mutually exclusive subsets according to a subsampling rate γ. Each subset is used to train one discriminator and the generator randomly queries one discriminator for one update.

Experimental Setup

In the experiments, we investigated the utility of the statistical test at different levels of privacy ε. For the DP-MW U test and all DP-synthetic data generation methods, except the DP GAN, we used ε values of 0.01, 0.1, 1, 5, and 10. For the DP GAN experiments, the ε values were 1, 2, 3, 4, 5, and 10. The higher minimum of $\varepsilon = 1$ was set due to differences between the DP GAN and the other methods. Every experiment was repeated 1,000 times, and the proportions of Type I and Type II errors were computed and evaluated at a significance level $\alpha = 0.05$.

Setup for Gaussian Data

In our experiments on Gaussian data using the DP-MW U test, DP Perturbed Histogram, Private-PGM, and MWEM, each method was applied to original dataset sizes of 50, 100, 500, 1,000, and 20,000 with a group ratio of 50% and at the different values of ε. In these methods, the original dataset size was considered to be of public knowledge, thus, the size of the generated DP-synthetic dataset was around or equal to the original size.

Experiments with DP Smoothed Histogram were performed by randomly sampling original Gaussian dataset of large size (i.e., dataset size of 20,000 with a group ratio of 50%). Then, the method was applied using the different values of ε, and for every ε synthetic data of size 50, 100, 500, and 1,000 were generated using the noisy probabilities returned by the method.

In all experiments with the GAN discriminator networks, a subsampling rate γ of 1/500 was used, resulting in mutually exclusive subsets of size 40. The sample size for the GAN training data was 20,000 in all settings and 1,000 different generators were trained with models saved at the chosen values of ε (1, 2,3, 4, 5, and 10). Five synthetic datasets of sizes 50, 100, 500, and 1,000 were sampled from each of the generators and MW-U tests were conducted on each of these synthetic datasets separately. The DP-hyperparameters were all set to $C = 1$ for the gradient clipping bound and 1.07 for the noise multiplier, following Chen et al.[25]

A summary of the settings for the experiments with original Gaussian data is provided in ►Table 1.

Setup for Real-World Data

The size of the PCa dataset constrained some of the experiments. Therefore, DP Smoothed Histogram and DP GAN experiments with this dataset were excluded, as these methods require a larger original dataset size (i.e., thousands of observations) to apply them accurately. On the other hand, the cardiovascular dataset size allowed us to carry out experiments with all the DP methods.

In the experiments with the PCa dataset, we applied each considered DP method at each epsilon value 1,000 times. While in the cardiovascular dataset experiments, we used the data to sample 1,000 original datasets for each dataset size: 50, 100, 500, 1,000, and 20,000; then, for each sampled dataset, we applied the DP methods at each epsilon. The proportion of Type II error was measured over the 1,000 repetitions for each experiment setting. For DP Smoothed Histogram and DP GAN, due to their nature, the experiments were performed differently; however, they had a similar setting to the ones with Gaussian signal data.

Setup for Simulated Multivariate Data

In the experiments with a simulated multivariate dataset, we considered the Private-PGM and MWEM synthesizers. Using the generated DP-synthetic data, we empirically assessed the proportion of Type I and Type II errors for the MW U test for an ordinal variable (PI-RADS score), Student's *t*-test for a normally distributed continuous variable (age), median test for another continuous variable (PSA), and chi-squared test of independence for a binary variable (use of 5-ARIs medication).

For these experiments, we generated 1,000 simulated multivariate datasets for dataset sizes of 50, 100, 500, 1,000, and 20,000. Subsequently, for each simulated dataset, we generated DP-synthetic data of the same size at each epsilon value. The proportions of Type I and Type II errors were measured across the DP-synthetic datasets, with the condition that the requirements for running the statistical

Table 1 Setup for experiments using original Gaussian data

DP method	Original dataset size, group ratio 50%	DP-synthetic dataset size	Privacy budget
DP-MW U test	50, 100, 500, 1,000, 20,000	N/A	$\varepsilon = 0.01, 0.1, 1, 5, 10$
DP Perturbed Histogram Private-PGM MWEM	50, 100, 500, 1,000, 20,000	Similar to the original dataset	$\varepsilon = 0.01, 0.1, 1, 5, 10$
DP Smoothed Histogram	20,000	50, 100, 500, 1,000	$\varepsilon = 0.01, 0.1, 1, 5, 10$
DP GAN	20,000	50, 100, 500, 1,000	$\varepsilon = 1, 2, 3, 4, 5, 10$

Note: For the DP-MW U test, DP-synthetic dataset size is not applicable ("N/A"), because this method is computed on the original sensitive data.

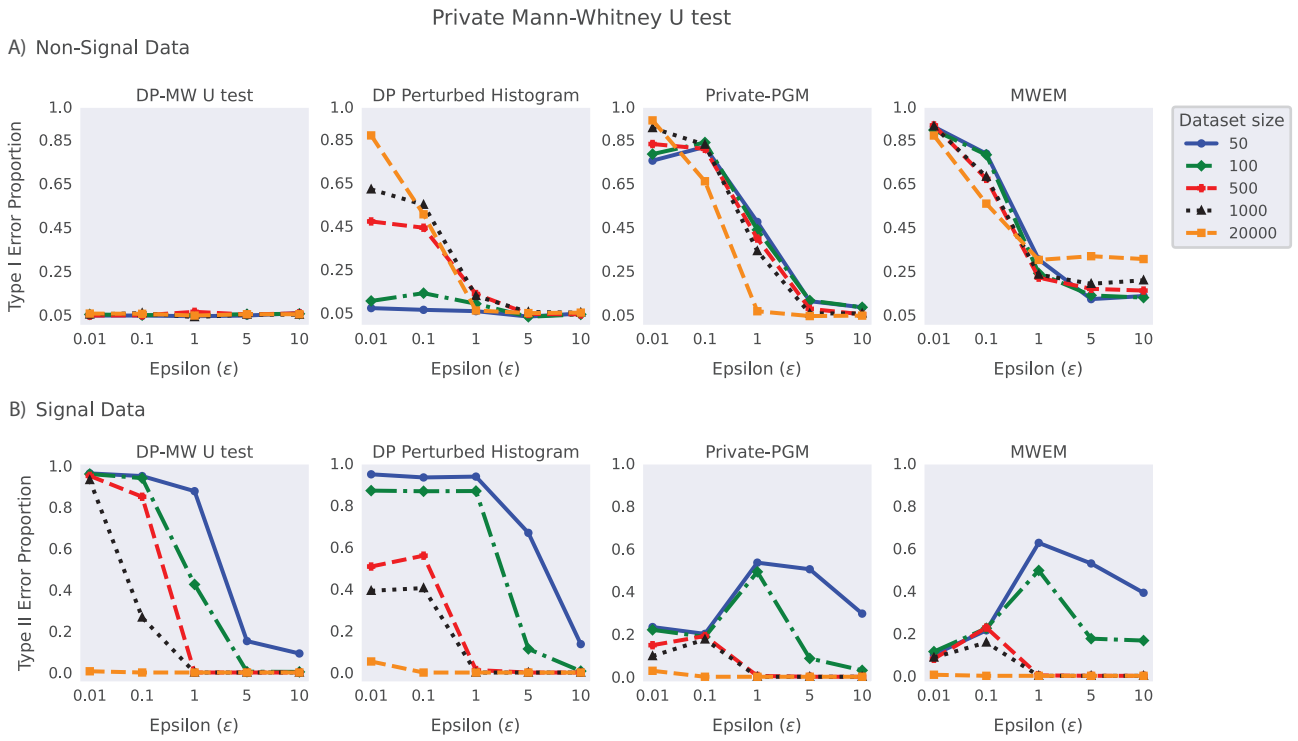

Fig. 4 The proportion of Type I and Type II errors for the Mann–Whitney U test using four differentially private (DP) methods: DP-MW U test, DP Perturbed Histogram, Private-PGM, and MWEM at different privacy budget (ε). The dataset size indicates the size of the original data used in the experiments by the DP methods. The proportions of Type I error and Type II error were measured over 1,000 repetitions of the experiment using Gaussian (**A**) non-signal data and (**B**) signal data, respectively.

test were met in at least 50 of the generated DP-synthetic datasets (see ►**Supplementary Material A.2** for further details on cases when tests are undefined, such as when the DP-synthetic data consists of only single class).

Results

Gaussian Data

In ►**Fig. 4A**, experiments on Gaussian non-signal data (i.e., both groups originate from the same Gaussian distribution) show that when the DP-MW U test is applied to the 1,000 datasets, the proportion of Type I stays close to $\alpha = 0.05$ for all dataset sizes at all ε. Meanwhile, the MW U test on DP-synthetic data from DP Perturbed Histogram, Private-PGM, and MWEM has a high proportion of Type I error for $\varepsilon < 5$, falsely indicating a significant difference between the two groups. From these DP methods, DP Perturbed Histogram and Private-PGM benefit of having a large original dataset size (i.e., 20,000), as ε can be reduced to 1 while still having a Type I error close to $\alpha = 0.05$. MWEM is the method with the worst performance as the proportion of Type I error for all sample sizes stays above 0.05 even for $\varepsilon = 10$.

►**Fig. 4B** presents the results for Gaussian signal data where a difference between the two groups exists (i.e., normally distributed data of two groups with means 1 standard deviation apart). From these results, we observed that the MW U test Type II error for all the DP methods, with low ε, can be reduced by increasing the dataset size, corroborating the trade-off that exists between privacy, utility, and dataset size.

Results for the MW U test on DP-synthetic data from DP Smoothed Histogram and DP GAN are presented in ►**Fig. 5**. The DP Smoothed Histogram method controls the Type I error reliably. However, the price for this is that in most of our experiment settings, it has high Type II error, meaning that the real difference between the groups present in the original data is lost in the DP-synthetic data generation process. DP GAN shows very high Type I error that as an interesting contrast to the other methods grows as privacy level is reduced.

To summarize, these results show that except for DP Smoothed Histogram, all the DP-synthetic data generation methods have highly inflated Type I error. This means that they are prone to generating data from which false discoveries are likely to be made. For the histogram-based methods, increased Type I error was associated with increased level of privacy, the effect being especially clear for $\varepsilon < 5$. ►**Fig. 6** presents an example of false discovery on synthetic data generated with the DP Perturbed histogram at $\varepsilon = 0.1$, and also demonstrates how DP Smoothed Histogram does not exhibit the same behavior.

Real-World Data

►**Fig. 7A** shows the results of experiments conducted with the PCa dataset. The DP-MW U test performs as expected for an original dataset size of 500 with a group ratio of approximately 50%. The null hypothesis is rejected for $\varepsilon \geq 1$, while for $\varepsilon < 1$ it is often not rejected. Similar behavior is present in DP-synthetic data from DP Perturbed Histogram and MWEM, yet

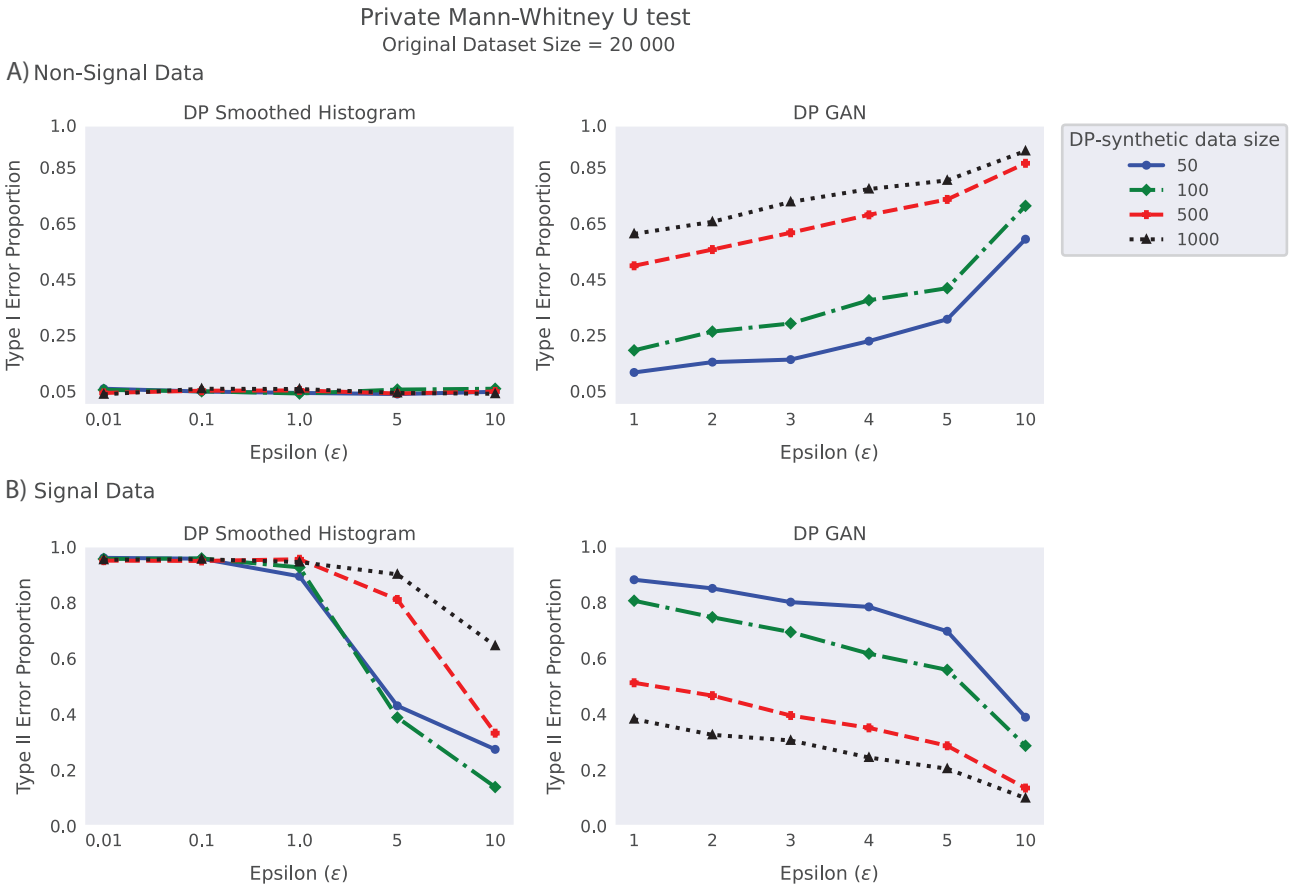

Fig. 5 The proportion of Type I and Type II errors of MW U test applied to synthetic data generated from DP Smoothed Histogram and DP GAN. The size of the original dataset is 20,000 with a group ratio of 50%. DP-synthetic data of sizes 50, 100, 500, and 1,000 were generated from both methods. The proportions of Type I error and Type II error were measured over 1,000 DP-synthetic datasets using Gaussian (**A**) non-signal data and (**B**) signal data, respectively. DP, differentially private.

the chance of rejecting the null hypothesis when $\varepsilon < 1$ is higher than in the DP-MW U test. In DP-synthetic data from Private-PGM, the null hypothesis is rejected for $\varepsilon \geq 5$ more often than for $\varepsilon < 5$.

The experiment results for the DP-MW U test, DP Perturbed Histogram, Private-PGM, and MWEM applied to the Cardiovascular Disease dataset are presented in ►**Fig. 7B**. In this dataset, we observe that MWEM and Private-PGM are the methods that benefit the most from increasing the original sample size, as stronger privacy guarantees can be provided without the MW U test losing power. These results agree with the ones obtained when using Gaussian signal data.

Results for DP Smoothed Histogram and DP GAN applied to the cardiovascular dataset are presented in ►**Fig. 8**. With DP Smoothed Histogram, Type II error is on an acceptable level when $\varepsilon \geq 5$ and the sample size is 500 or 1,000, whereas for lower ε values the effect is not found. DP GAN results have lower Type II error, but given how high Type I error the method shows in the non-signal experiments, the approach is less reliable compared to the DP Smoothed Histogram method.

Simulated Multivariate Data

In ►**Fig. 9**, the proportion of Type I errors for various statistical tests (i.e., MW U test, Student's *t*-test, median test, and chi-squared test) is presented. From these results, we observe that false discoveries are also prone to occur similarly to the previous experiments with only two variables. The validity of the tests is preserved only for largest tested privacy budgets combined with large amounts of the original sensitive data. Same kind of trend was observed for all statistical tests under consideration. For Private-PGM, a substantial drop in Type I error was observed for $\varepsilon = 0.01$ and dataset size <20,000. On closer examination, we observed that with the smallest privacy budgets, the size of the smaller of the two groups tends to be very small or even zero. This can be seen from the numbers of times the test requirements failed, as presented in ►**Supplementary Material A.2**, where the tests fail when the size of smaller group is zero. The power of all evaluated tests strongly depends on the group size imbalance in the sample, so that for a fixed sample size they have the highest power for equal group sizes and the power shrinks to zero when the smaller group size goes to zero. Therefore, the tendency of the low privacy budgets to produce imbalanced samples counters the tendency to produce fake group differences to some extent.

In the case of Type II error proportions (►**Fig. 10**), the results depend on the magnitude of group differences in the original data, how it is preserved by the GaussianCopula-Synthesizer, and the size of the simulated dataset. As a

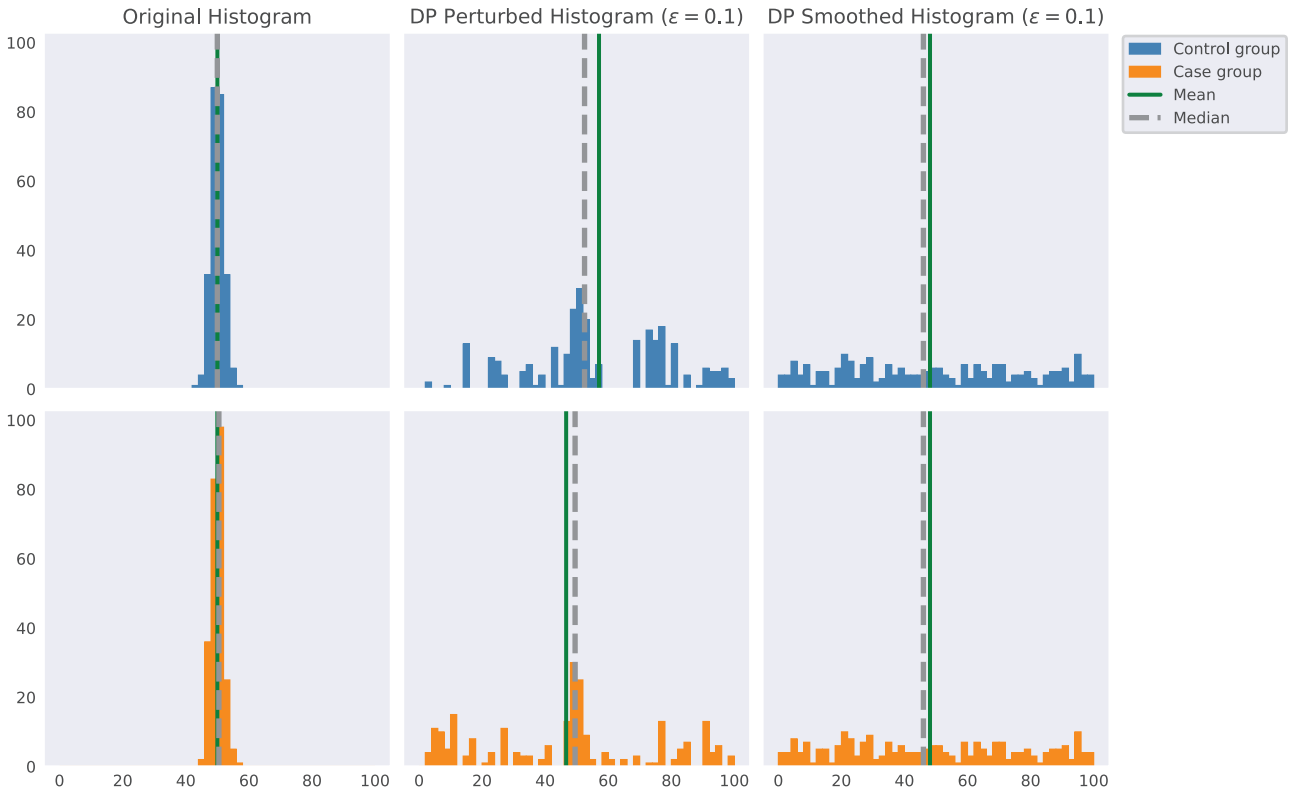

Fig. 6 Example of the two groups' distributions in a non-signal original dataset of size 500 (U stat = 31,460.5, *p*-value = 0.8953) and the corresponding distributions for synthetic data generated using DP Perturbed Histogram (MW U stat = 38,191.5, *p*-value = 0.00001774) and DP Smoothed Histogram (MW U stat = 29,621.5, *p*-value = 0.3314) with $\varepsilon = 0.1$ as the privacy budget. With such high level of privacy enforced neither of the DP-synthetic datasets preserves well the structure of the original data. The DP Perturbed Histogram has the tendency to create artificial differences between the two groups such that result in low *p*-values for MW U test, whereas with DP Smoothed Histogram method both the generated case and control groups follow similar close to uniform distributions. DP, differentially private.

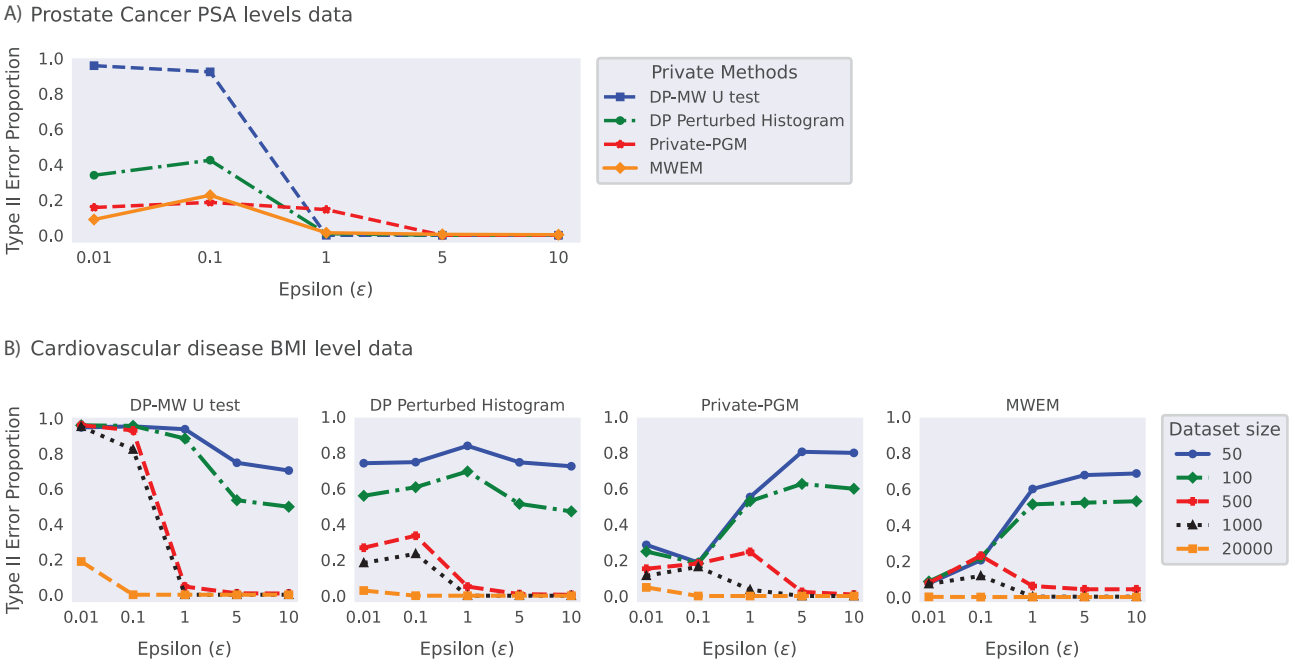

Fig. 7 The proportion of Type II error for the Mann–Whitney U test using four DP methods: DP-MW U test, DP Perturbed Histogram, Private-PGM, and MWEM applied to (**A**) the PSA level data in the prostate cancer dataset (dataset size = 500) and (**B**) the body mass index (BMI) data in the Kaggle Cardiovascular Disease dataset. DP, differentially private.

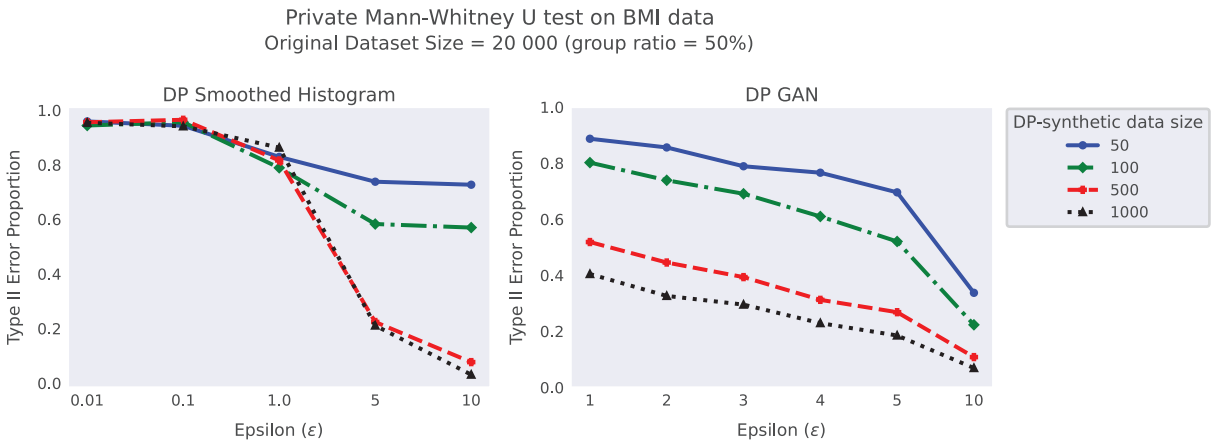

Fig. 8 The proportion of Type II error of MW U test applied to synthetic data generated from DP Smoothed Histogram and DP GAN. The original dataset of size 20,000 with a group ratio of 50% was drawn from the publicly available Cardiovascular Disease dataset. DP-synthetic data of sizes 50, 100, 500, and 1,000 were generated using both DP methods 1,000 times. DP, differentially private.

baseline or point of reference, we first present the Type II error probabilities computed over the 1,000 simulated multivariate datasets that represent the original sensitive data before the synthetic data are generated based on them. Then, we illustrate the corresponding Type II error probabilities for the DP-synthetic data with different privacy budgets. For the synthetic data, especially for $\varepsilon < 10$, we observe that the Type II errors are often lower than those of on the original data, indicating that true group differences are discovered more often from the synthetic data than from the original. However, this is explained perfectly by the large Type I error probabilities presented in ►Fig. 9, indicating that the fake group differences present in the synthetic data are so strong that they end up getting discovered rather than the true ones that are too weak to be discovered from the original data. For $\varepsilon = 0.01$, the large Type II error of Private-PGM also mirrors the low Type I error, caused by the loss of power due to the group size imbalance.

Discussion

This study investigated to what extent the validity and power of independent sample tests are preserved in DP-synthetic data. Experimental results on Gaussian, real-world, and multivariate simulated data demonstrate that the generated DP-synthetic data, especially with strong privacy guarantees ($\varepsilon \leq 1$), can lead to false discoveries. We empirically show that many state-of-the-art DP methods for generating synthetic data have highly inflated Type I error when the privacy level is high. These results indicate that false discoveries or inferences are likely to be drawn from the DP-synthetic data produced by these DP methods. Our findings are in line with other studies that have presented or stated that DP-synthetic data can be invalid for statistical inference and indicated the need for methods that are noise-aware in order to produce accurate statistical inferences.[17,66–69]

Additionally, it is necessary to be cautious when analyzing Type II error results, as this is only meaningful for valid tests where the Type I error is properly controlled. The Type II error tends to decrease with the increase of Type I error, as these errors are inversely related. In our study, the only DP method based on synthetic data generation that had a valid Type I error over all the privacy budgets tested was the DP Smooth Histogram method. However, the method is applicable only when the original dataset size is fairly large (e.g., $n = 20{,}000$ in our experiments) and tended to have high Type II error when the amount of privacy enforced was high (e.g., $\varepsilon \leq 1$). For DP Perturbed Histogram and Private-PGM methods, both Type I and Type II errors remained low for $\varepsilon \geq 5$, whereas MWEM and DP GAN did not provide valid Type I error levels even with lowest privacy values tested.

The main advantage of releasing DP-synthetic data, as opposed to releasing only analysis results from the original data, is that it can be ideally used to support a wide range of analyses by different users. Due to postprocessing property of DP, any type or number of analyses done on the synthetic data are also guaranteed to be DP with no further privacy budget needed. However, if the only goal is to perform a limited number of predefined analyses, it makes more sense to do these on the original data with DP methods. This is illustrated in our experiments by the DP-MW U test baseline that always outperforms analyses done on DP-synthetic data. As a middle ground between these approaches, an active area of research is to develop such DP synthetization methods where the data are optimized to support certain types of analyses well, such as PrivPfC[70] for classifier training and various Bayesian noise-aware DP synthetic data generation methods.[69]

There are limitations in our study that could be addressed in future research. One limitation is that marginal- or histogram-based DP methods require continuous variables to be discretized. This discretization must be performed in a private manner or based on literature to avoid leaking private information. Besides, it is well known that the number of bins used to discretize the data has a significant impact on the quality of the resulting data.[16,43] Therefore, choosing the number of bins is problem- and data-dependent and can affect the results. In our experiments with Gaussian data, the

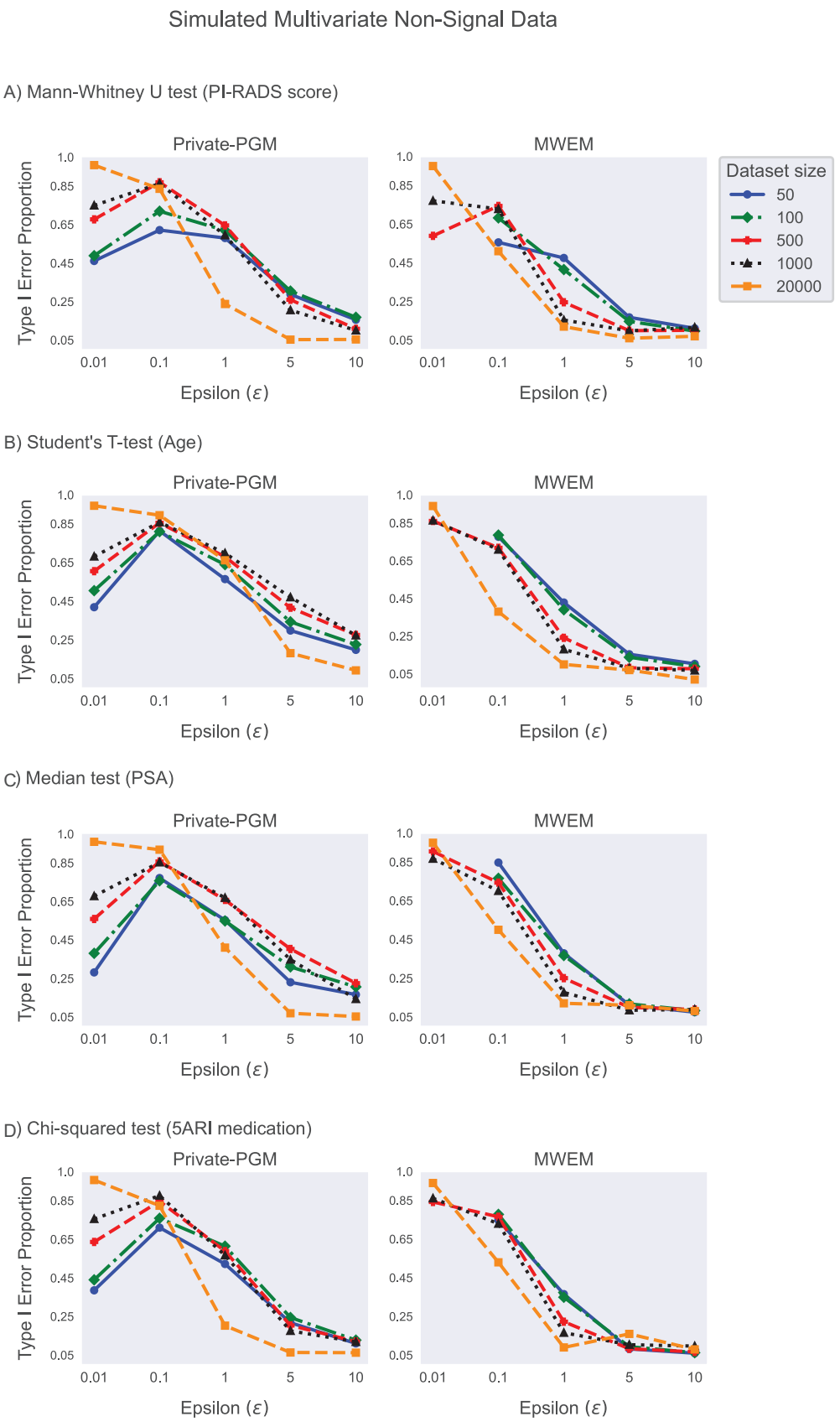

Fig. 9 Proportion of Type I error conditioned to the number of DP-synthetic datasets where the statistical test is feasible; (**A**) MW U test applied to an ordinal variable (PI-RADS score); (**B**) Student's *t*-test on a normally distributed variable (age); (**C**) median test on continuous variable (PSA); **(D)** Chi-squared test on a binary variable (5-ARI medication). DP, differentially private.

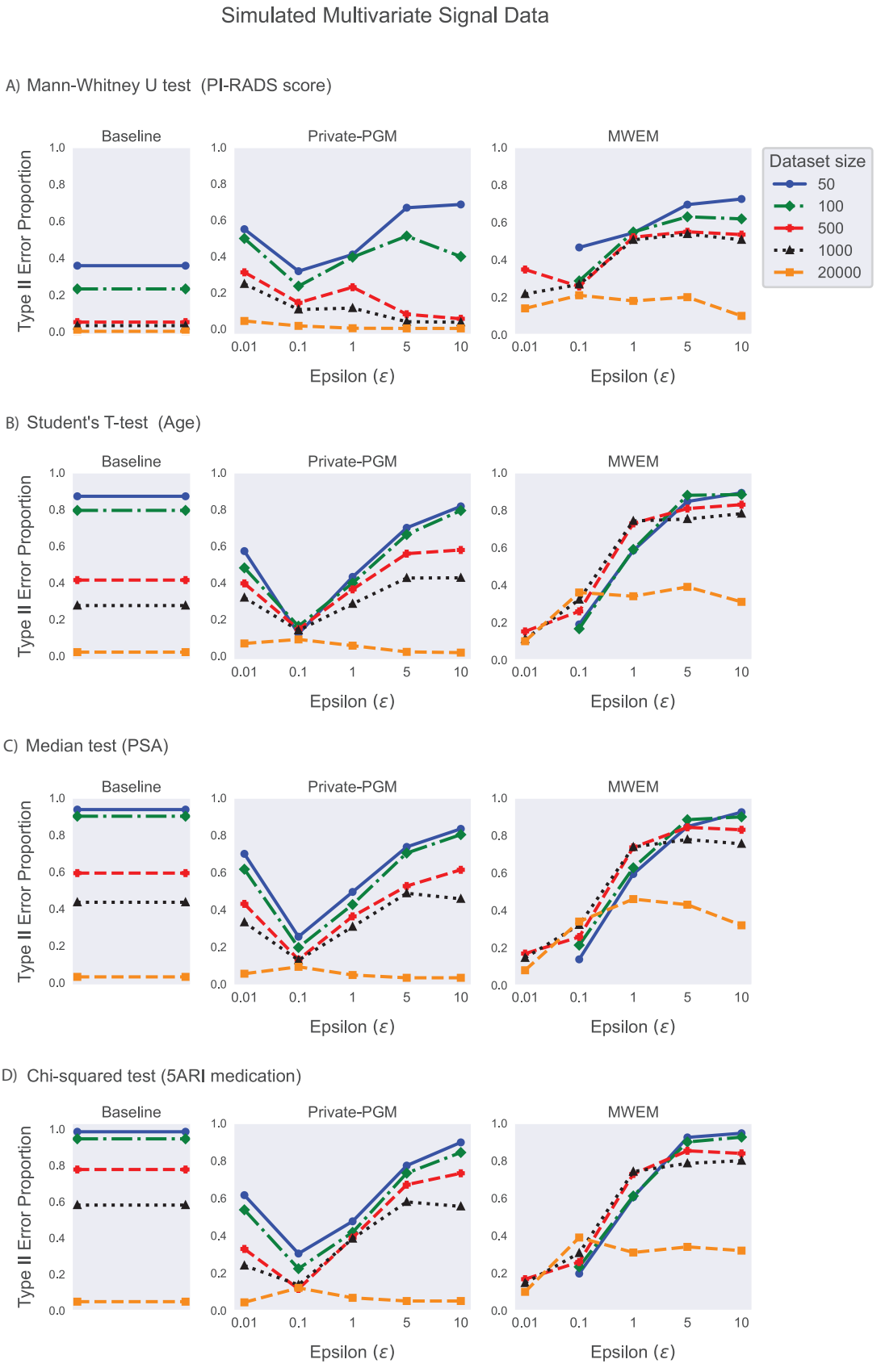

Fig. 10 Proportion of Type II error conditioned to the number of DP-synthetic datasets where the statistical test is feasible; (**A**) MW U test applied to an ordinal variable (PI-RADS score); (**B**) Student's *t*-test on a normally distributed variable (age); (**C**) median test on continuous variable (PSA); (**D**) chi-squared test on a binary variable (5-ARI medication). DP, differentially private.

continuous values were discretized using 100 bins. This number of bins was selected to show a possible extreme case where having bins empty or with small counts deteriorates the quality of the generated DP-synthetic data. On the other hand, for our experiments with real-world and multivariate simulated data, the number of bins used was determined based on domain knowledge and literature. Finally, testing different hyperparameter values for the DP method implementations could yield different results for the methods.

Conclusion

Our results suggest caution when releasing DP-synthetic data, as false discoveries or loss of information is likely to happen especially when a high level of privacy is enforced. To an extent, these issues may be mitigated by having large enough original datasets, selecting methods that are less prone to adding false signal to data, and by carefully comparing the quality of the DP-synthetic data to the original one based on various quality metrics (see, e.g., Hernadez et al[4]) before data release. Still, with current methods, DP-synthetic data may be a poor substitute for real data when performing statistical hypothesis testing, as one cannot be sure if the results obtained are based on trends that hold true in the real data, or due to artefacts introduced when synthetizing the data.

Funding

This work has received funding from Business Finland (grant number 37428/31/2020) and European Union's Horizon Europe research and innovation programme (grant number 101095384). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

Conflict of Interest

None declared.

Acknowledgement

The authors would like to express their gratitude to Peter J. Boström, Ivan Jambor, and collaborators for their support and contribution in providing the PCa datasets used in the real-world data experiments. We also thank Katariina Perkonoja for her insightful feedback regarding the experimental setup and the statistical tests, as well as the anonymous reviewers for their valuable comments.

References

1 El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ 2015;350(01):h1139
2 Rubin DB. Statistical disclosure limitation. J Off Stat 1993;9(02): 461–468
3 Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng 2021;5(06):493–497
4 Hernadez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. Methods Inf Med 2023;62 (S 01):e19–e38
5 Jordon J, Szpruch L, Houssiau F, et al. Synthetic data-what, why and how? arXiv preprint 2022. Accessed May 17, 2023 at: http://arxiv.org/abs/2205.03257
6 Chen D, Yu N, Zhang Y, Fritz M. GAN-Leaks: a taxonomy of membership inference attacks against generative models. Paper presented at: Proceedings of the ACM Conference on Computer and Communications Security. Virtual event, United States: ACM; 2020:343–362
7 Hayes J, Melis L, Danezis G, De Cristofaro E. LOGAN: membership inference attacks against generative models. arXiv preprint 2018. Accessed May 22, 2023 at: https://arxiv.org/abs/1705.07663v4
8 Stadler T, Oprisanu B, Troncoso C. Synthetic data–anonymisation groundhog day. arXiv preprint 2022. Accessed May 9, 2023 at: https://arxiv.org/abs/2011.07018
9 Carlini N, Brain G, Liu C, Erlingsson Ú, Kos J, Song D. The secret sharer: evaluating and testing unintended memorization in neural networks. Paper presented at: 28th USENIX Security Symposium (USENIX Security 19), Santa Clara, California, United States; 2019:267–284. Accessed May 17, 2023 at: https://www.usenix.org/conference/usenixsecurity19/presentation/carlini
10 Karr AF, Kohnen CN, Oganian A, Reiter JP, Sanil AP. A framework for evaluating the utility of data altered to protect confidentiality. Am Stat 2006;60(03):224–232
11 Boedihardjo M, Strohmer T, Vershynin R. Covariance's loss is privacy's gain: computationally efficient, private and accurate synthetic data. Found Comput Math 2024;24(01):179–226
12 Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T, eds. Theory of Cryptography Conference. Berlin: Springer Berlin Heidelberg; 2006:265–284
13 Dwork C, Roth A. The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci 2014;9(3–4):211–487
14 Wasserman L, Zhou S. A statistical framework for differential privacy. J Am Stat Assoc 2010;105(489):375–389
15 Gong M, Xie Y, Pan K, Feng K, Qin AK. A survey on differentially private machine learning. IEEE Comput Intell Mag 2020;15(02): 49–64
16 Xu J, Zhang Z, Xiao X, Yang Y, Yu G, Winslett M. Differentially private histogram publication. VLDB J 2013;22(06):797–822
17 Gaboardi M, Lim HW, Rogers R, Vadhan SP. Differentially private chi-squared hypothesis testing: goodness of fit and independence testing. Paper presented at: Proceedings of the 33rd International Conference on Machine Learning, New York, United States. PMLR; 2016:2111–2120
18 Task C, Clifton C. Differentially private significance testing on paired-sample data. Paper presented at: 16th SIAM International Conference on Data Mining, Miami, Florida, United States, May 5–7, 2016; SDM; 2016:153–161
19 Couch S, Kazan Z, Shi K, Bray A, Groce A. Differentially private nonparametric hypothesis testing. Paper presented at: Proceedings of the ACM Conference on Computer and Communications Security. ACM; 2019:737–751
20 Ferrando C, Wang S, Sheldon D. Parametric bootstrap for differentially private confidence intervals. arXiv preprint 2021. Accessed February 5, 2023 at: https://arxiv.org/abs/2006.07749
21 Chaudhuri K, Monteleoni C, Sarwate AD. Differentially private empirical risk minimization. J Mach Learn Res 2011; 12:1069–1109
22 Hardt M, Ligett K, McSherry F. A simple and practical algorithm for differentially private data release. Adv Neural Inf Process Syst 2012;3:2339–2347
23 Ping H, Stoyanovich J, Howe B. DataSynthesizer: privacy-preserving synthetic datasets. Paper presented at: 29th International

Conference on Scientific and Statistical Database Management; June 27, 2017; Chicago, Illinois, United States. ACM; 2017:1–5
24 Snoke J, Slavković A pMSE mechanism: differentially private synthetic data with maximal distributional similarity. arXiv preprint 2018. Accessed October 5, 2022 at: https://arxiv.org/abs/1805.09392
25 Chen D, Orekondy T, Fritz M. GS-WGAN: a gradient-sanitized approach for learning differentially private generators. Adv Neural Inf Process Syst 2020;33:12673–12684
26 McKenna R, Miklau G, Sheldon D. Winning the NIST Contest: a scalable and general approach to differentially private synthetic data. J Priv Confid 2021;11(03):10.29012/jpc.778
27 Nachar N. The Mann-Whitney U: a test for assessing whether two independent samples come from the same distribution. Tutor Quant Methods Psychol 2008;4(01):13–20
28 Zar JH. Biostatistical Analysis. 5th ed. New Jersey: Pearson Prentice Hall; 2010
29 Okeh UM. Statistical analysis of the application of Wilcoxon and Mann-Whitney U test in medical research studies. Biotechnol Mol Biol Rev 2009;4(06):128–131
30 Fay MP, Proschan MA. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat Surv 2010;4:1–39
31 Kim TK. T test as a parametric statistic. Korean J Anesthesiol 2015; 68(06):540–546
32 Conover WJ. Practical Nonparametric Statistics. New York, NY: Wiley; 1999
33 McHugh ML. The chi-square test of independence. Biochem Med (Zagreb) 2013;23(02):143–149
34 Casella G, Berger RL. Statistical inference. 2nd ed. Pacific Grove, CA, USA: Duxbury Press; 2002
35 Arnold C, Neunhoeffer M. Really useful synthetic data–a framework to evaluate the quality of differentially private synthetic data. arXiv preprint 2020. Accessed May 19, 2023 at: https://arxiv.org/abs/2004.07740v2
36 Abadi M, McMahan HB, Chu A, et al. Deep learning with differential privacy. Paper presented at: 2016 ACM SIGSAC Conference on Computer and Communications Security; October 24, 2016; Vienna, Austria. ACM; 2016:308–318
37 Abay NC, Zhou Y, Kantarcioglu M, Thuraisingham B, Sweeney L. Privacy preserving synthetic data release using deep learning. Paper presented at: European Conference on Machine Learning and Knowledge Discovery in Databases; September 10, 2018; Dublin, Ireland. Springer; 2018:510–526
38 Jordon J, Yoon J, Van Der Schaar M. PATE-GAN: Generating synthetic data with differential privacy guarantees. Paper presented at: International Conference on Learning Representations; May 6, 2019; New Orleans, Louisiana. ICLR; 2019
39 Bowen CM, Snoke J. Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge. J Priv Confid 2021;11(01):10.29012/jpc.748
40 Bowen CM, Liu F. Comparative study of differentially private data synthesis methods. Stat Sci 2020;35(02):280–307
41 Cai K, Lei X, Wei J, Xiao X. Data synthesis via differentially private markov random fields. Proc VLDB Endow 2021;14(11):2190–2202
42 Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X. Privbayes: Private data release via bayesian networks. ACM Trans Database Syst (TODS) 2017;42(04):1–41
43 McKenna R, Mullins B, Sheldon D, Miklau G. AIM: an adaptive and iterative mechanism for differentially private synthetic data. Proc VLDB Endow 2022;15(11):2599–2612
44 Wang T, Yang X, Ren X, Yu W, Yang S. Locally private high-dimensional crowdsourced data release based on copula functions. IEEE Trans Serv Comput 2022;15(02):778–792
45 Ren X, Yu CM, Yu W, et al. LoPub: high-dimensional crowdsourced data publication with local differential privacy. IEEE Trans Inf Forensics Security 2018;13(09):2151–2166
46 Chen R, Li H, Qin AK, Kasiviswanathan SP, Jin H. Private spatial data aggregation in the local setting. Paper presented at: 2016 IEEE 32nd International Conference on Data Engineering, ICDE 2016; 2016:289–300
47 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Commun ACM 2020;63(11):139–144
48 Ian Goodfellow. NIPS 2016 tutorial: generative adversarial networks. arXiv preprint 2016. Accessed May 19, 2023 at: https://arxiv.org/abs/1701.00160v4
49 Wilcoxon F. Individual comparisons by ranking methods. Biom Bull 1945;1(06):80–83
50 Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947;18(01):50–60
51 Freidlin B, Gastwirth JL. Should the median test be retired from general use? Am Stat 2010;54(03):161–164
52 Jambor I, Boström PJ, Taimen P, et al. Novel biparametric MRI and targeted biopsy improves risk stratification in men with a clinical suspicion of prostate cancer (IMPROD Trial). J Magn Reson Imaging 2017;46(04):1089–1095
53 Jambor I, Verho J, Ettala O, et al. Validation of IMPROD biparametric MRI in men with clinically suspected prostate cancer: a prospective multi-institutional trial. PLoS Med 2019;16(06):e1002813
54 Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, Redwine E. Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. N Engl J Med 1987;317(15):909–916
55 Catalona WJ, Smith DS, Ratliff TL, et al. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. N Engl J Med 1991;324(17):1156–1161
56 Ulianova S. Cardiovascular Disease dataset | Kaggle. 2019. Accessed October 12, 2022 at: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
57 Larsson SC, Bäck M, Rees JMB, Mason AM, Burgess S. Body mass index and body composition in relation to 14 cardiovascular conditions in UK Biobank: a Mendelian randomization study. Eur Heart J 2020;41(02):221–226
58 Patki N, Wedge R, Veeramachaneni K. The synthetic data vault. Paper presented at: Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016; 2016:399–410
59 Virtanen P, Gommers R, Oliphant TE, et al; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17(03):261–272
60 Canonne CL, Kamath G, Steinke T. The discrete Gaussian for differential privacy. J Priv Confid 2022;12(01):10.29012/jpc.784
61 McKenna R, Miklau G, Sheldon D. Private-PGM. GitHub 2021. Accessed April 8, 2022 at: https://github.com/ryan112358/private-pgm
62 Hardt M, Ligett K, McSherry F. Private Multiplicative Weights (MWEM). GitHub 2020. Accessed November 7, 2022 at: https://github.com/mrtzh/PrivateMultiplicativeWeights.jl
63 Chen D. GS-WGAN. GitHub 2020. Accessed October 8, 2022 at: https://github.com/DingfanChen/GS-WGAN
64 Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. Paper presented at: Proceedings of the 33rd International Conference on Neural Information Processing Systems; December 8, 2019; Vancouver, Canada. Curran Associates Inc.; 2019:8026–8037
65 Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. Paper presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; December 4, 2017; Long Beach, California. Curran Associates Inc.; 2017:5769–5779
66 Charest A-S. How can we analyze differentially-private synthetic datasets? J Priv Confid 2011;2(02):21–33
67 Charest A-S. Empirical evaluation of statistical inference from differentially-private contingency tables. Paper presented

at: International Conference on Privacy in Statistical Databases; September 26, 2012; Palermo, Italy. Springer-Verlag; 2012: 257–272

68 Giles O, Hosseini K, Mingas G, et al. Faking feature importance: a cautionary tale on the use of differentially-private synthetic data. arXiv preprint 2022. Accessed May 19, 2023 at: https://arxiv.org/abs/2203.01363

69 Räisä O, Jälkö J, Kaski S, Honkela A. Noise-aware statistical inference with differentially private synthetic data. Paper presented at: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics; April 25, 2023; Valencia, Spain. PMLR; 2023:3620–3643

70 Su D, Cao J, Li N, Lyu M. PrivPfC: differentially private data publication for classification. VLDB J 2018;27(02):201–223

