PRIVACY-PRESERVING SYNTHETIC DATA FOR THE CCHIC HEALTHCARE DATASET

ABSTRACT

The proliferation of individual level datasets has opened up new research opportunities. However, the privacy of individual information is often tightly restricted. This creates difficulties in obtaining and using the data to deliver answers to key research or policy questions. Methods exist for creating synthetic populations that are representative of the original data. However, it is unclear to what extent these methods can act as an alternative to more traditional anonymisation methods in real-world settings with strict privacy requirements (e.g., in healthcare). Here, we assess the applicability of privacy-preserving generative models for releasing synthetic versions of the non-public CCHIC (Critical Care Health Informatics Collaborative) dataset, which contains data from patients in intensive care units in the UK. We present an empirical analysis which applies PrivBayes, a differentially-private Bayesian generative model, to a subset of variables from CCHIC. We assess the privacy of the generated synthetic datasets by running two types of motivated intruder attacks, membership inference and attribute inference attacks, and measuring their likelihood of success; we compare this to two baselines (traditionally anonymised datasets and random/prior guess). We do the same comparison for various utility measures. This initial investigation shows that synthetic data offer better protection than anonymised data against attribute inference attacks and similar protection against membership inference attacks and that intruders gain only a marginal advantage when having access to synthetic data versus access only to prior information. The trade-off between privacy and utility also favours synthetic data although some distributional distortion should be expected even with low privacy protection.

UCLH has received approval from the CCHIC governing body to evaluate the existing anonymisation processes and other data publishing approaches, such as privacy-preserving generative models and synthetic datasets. This report documents the steps that the Alan Turing Institute and UCLH teams took in this project. This work was performed under the authority and supervision of the leads for the NIHR HIC Critical Care theme (UCLH) at all times.

1 Introduction

Access to high-quality individual-level data is essential for many data science, machine learning (ML) and other research projects in the public, private and academic sectors. However, sharing data outside of organisations that collect them is often difficult as there is a risk to the privacy of individuals and communities whose data have been collected and stored. Being able to share data with a wide number of researchers and organisations is crucial to promoting research in various areas and extracting insights that can help influence decision and policy making.

Healthcare data is of particular interest where a need exists for high-quality data, and the problem of data sharing is particularly acute. Dataset access is typically tightly restricted, for example, to specific users within Trusted Research Environment service, such as Data Safe Haven (DSH) environments. This creates difficulties in obtaining and using the data as well as difficulties in working openly and reproducibly since full analyses cannot then be shared. Being able to share electronic health records, clinical trials or registry data for entire populations more widely could drive life-saving insights and make new and powerful connections that we are currently unable to see.

One common approach for addressing the problem of wider data sharing is to use anonymisation, in which various transformations are applied to remove sensitive information before releasing the data. Examples for such transformations include removing direct identifiers, removing outliers and micro-aggregation. The aim is to guarantee a certain level of protection for individuals' privacy, for example, a certain level of k-anonymity or l-diversity in the dataset (see Section 4.1). The main advantage of anonymisation techniques is simplicity, both in terms of implementation and understanding or interpreting the privacy guarantees. However, the dimensionality and sparsity of realistic datasets make them vulnerable to privacy attacks, even when several steps of anonymisation have been applied to them. It has been shown that anonymisation techniques fail to protect the privacy of individuals in various settings [Emam et al., 2011, Henriksen-Bulmer and Jeary, 2016, Stadler et al., 2021]. Moreover, these traditional techniques can sometimes lead to datasets with low utility as they perturb the data in ways that cause important patterns to be lost [Stadler et al., 2021].

An alternative approach to address the shortcomings of traditional anonymisation techniques is to generate and release synthetic dataset. These datasets are designed to learn and replicate the statistical properties of the original data. They also aim to avoid revealing the identity of individuals in the original, sensitive dataset. Here, we focus on synthetic datasets generated by sampling from probabilistic generative models. Various privacy-preserving generative models have been proposed which range from Bayesian networks to generative deep learning approaches [Yoon et al., 2019a, Zhang et al., 2017a]. They often come with embedded differential privacy mechanisms that guarantee a level of privacy, normally measured by a privacy budget ϵ. These methods offer a theoretical guarantee on the level of privacy, rather than the more arbitrary definitions of privacy offered by k-anonymity and similar measures. The viability of synthetic data for use in real-world data releases is currently investigated in academic literature; there are mixed results on whether

synthetic methods can protect individual privacy while maintaining high utility (i.e., they allow useful analyses to be performed by not distorting the properties of the original data).

In this report, we aim to investigate this question in a realistic setting and assess the usability of synthetic data sharing as an alternative to traditional anonymisation techniques for a healthcare dataset. This report documents the outcomes of a collaborative project involving the Alan Turing Institute and UCLH's CCHIC (Critical Care Health Informatics Collaborative) team. CCHIC is a multi-centre intensive care database in the UK[1]. It records hundreds of fields including time-varying fields of patients during their stay in intensive care units across five NHS trusts in England. The recorded variables include patient demographics, time of admission and discharge, survival status, diagnosis, physiology, laboratory, nursing and drugs. These data can help gain invaluable insights into events and outcomes in ICU clinics and can help influence policy decisions.

At the moment, CCHIC has a pipeline[2] in place which is used to release subsets of their dataset to researchers after applying traditional anonymisation methods to the original data. CCHIC are looking for ways to improve their data release service, and synthetic methods are a promising option. It is worth mentioning that there has been increased public discussion on using synthetic data for healthcare[3], as well as several practical initiatives[4] during the last few years.

The following sections present our empirical analysis using a differentially-private Bayesian generative model called PrivBayes. We generate various synthetic datasets with different privacy levels for a subset of important and identifying variables in the CCHIC database. We assess the privacy of the generated datasets using two types of motivated intruder attacks: membership inference and attribute inference attacks. We calculate the probability of success for each of these attacks and compare it across privacy levels and against a number of baselines, including traditionally anonymised datasets and random/prior guess. We do the same comparison for various utility measures as measured by predictive accuracy of an ML model as well as means, medians and frequencies of the variables.

The codes developed for this work can be found in two repositories. One contains codes to extract data from the database on the UCL's Data Safe Haven environment[5] and apply some preprocessing steps, and the other one is a Python library to synthesise data and assess utility and privacy[6]. For our experiments we reused and extended code from the QUiPP (Quantifying Utility and Preserving Privacy in synthetic datasets) repository[7] as well as the repository developed as part of the "Synthetic Data - Anonymisation Groundhog Day" paper[8] by Stadler et al. [2021].

The main contributions of this work are as follows:

[1] https://discovery.ucl.ac.uk/id/eprint/10050778/

[2] https://github.com/CC-HIC/ccanonym

[3] https://bit.ly/3CArnYy

[4] e.g., https://odileeds.org/events/synae/; https://healthdatainsight.org.uk/project/the-simulacrum/; https://cprd.com/content/synthetic-data.

[5] https://github.com/alan-turing-institute/QUIPP-CC-HIC

[6] https://github.com/alan-turing-institute/synthetic_data_release

[7] https://github.com/alan-turing-institute/QUIPP-pipeline

[8] https://github.com/spring-epfl/synthetic_data_release

1. Understand if and how synthetic data can be used to make the CCHIC data more widely accessible to researchers, while maintaining high levels of privacy and utility.
2. Benchmark synthetic data against the non-public CCHIC dataset and the existing anonymised version of it.
3. Engage in a discussion with CCHIC domain experts and data owners to explore utility and privacy metrics which are applicable to synthetic data while being easy to communicate to both experts and the public, for example, probability of success for different types of intruder attacks as measures for privacy.
4. Inform the next stages of CCHIC's data release planning.
5. Transferring and extending methodology developed by the Alan Turing Institute (and especially the QUiPP project[9]) and third parties to a real-world healthcare dataset.
6. Create openly available code that will facilitate the creation of synthetic data for healthcare settings and the exploration of the privacy/utility trade-off by data publishing organisations and researchers.

2 Data overview

The CCHIC data contain the electronic health records related to episodes of critical care. In Section 2.1, we describe the original, non-public dataset. In 2016, an anonymised version of this dataset was released which we will describe in Section 2.2.

2.1 CCHIC dataset and selection of variables

We accessed the original data via Data Safe Haven (DSH), a platform for storing, handling and analysing identifiable data. DSH allows approved users to connect to a virtual machine within the environment and thus access and analyse the data. Data and code within the DSH need to be reviewed before leaving the system.

The CCHIC database contains data from ICUs in five NHS Trusts. Each patient admission is registered as an episode with demographic information, time of admission and discharge, survival status and other variables. Each episode is associated with a number of events which contain different measurements taken during the patient stay at the ICU, for example, clinical measurements and drug administration. Many of these events are time-series data. In total, there are 47,836 records at the time this analysis was conducted.

In this initial investigation, we did not use all the columns in the original database due to computational and complexity concerns. We instead chose a subset which includes the following columns (the list also shows their data categories as stated in the Standard Operating Procedure document[10]):

- age (key variable)

[9] https://github.com/alan-turing-institute/QUIPP-pipeline
[10] https://github.com/CC-HIC/ccanonym/blob/master/inst/SOP%20data%20release.pdf

- sex (key variable)
- height (key variable)
- weight (key variable)
- ethnicity (key variable)
- hour/week/month of admission (date-time variable)
- HIV/AIDS (sensitive)
- Biopsy proven cirrhosis (sensitive)
- "dead or alive on discharge" which is survival for a given ICU episode (non-identifying).

We chose this set of variables because we wanted:

- a subset that contains the main key variables that could be used to identify people and infer membership (excluding direct identifiers).
- to include a few sensitive variables that could be the target of attribute inference attempts.
- to compare the risk of releasing the HIV and Cirrhosis variables via synthesis versus anonymisation. These variables are currently not released in the anonymised version of CCHIC due to privacy concerns.
- to train a ML model that predicts an outcome ("dead or alive on discharge") from key/sensitive variables in order to assess utility for a basic research task.

The final dataset used in this analysis had 47,836 rows and 11 columns.

2.2 Anonymised dataset

An anonymised version of the CCHIC data was released in 2016. This dataset was assembled to enable clinical and health science research teams to interrogate these rich data streams with the aim of improving the care of future patients.

The anonymisation steps to protect the existing patients are documented in The Standard Operating Procedure (SOP) document in the "CC-HIC" repository.[11] Here, we briefly discuss some of the steps relevant to our experiments in Section 6. It should be noted that the released data is anonymised and not pseudonymised, where the latter term indicates that a unique ID is available that could be used to re-identify the data.[12]

The goal of anonymisation is to make the likelihood of re-identification low or non-existent. At the same time, there is a tradeoff between re-identification risk and information loss. Generally speaking, the utility of the data decreases as the risk of re-identification decreases. When creating an anonymised dataset, both of these measures (i.e., re-identification or disclosure risk and information loss) should be taken into account. In CCHIC, the anonymisation methodology consisted of several

[11] https://github.com/CC-HIC/ccanonym

[12] As stated in SOP, pseudonymisation is used for transferring data between organisations where data linkage must be undertaken, and where one of the organisations does not hold all the necessary permissions to handle those data.

steps, including removal of direct identifiers (e.g., NHS number or hospital number), removing living subjects wherever possible, change absolute date and time stamps to relative ones, removing high risk individuals and specific opt-outs (e.g., well-known public figures) and aggregating continuous and date-time key variables.

After this step, the remaining key and sensitive variables are processed according to the specified k-anonymity and l-diversity constraints. (In the anonymised version of CCHIC, the minimum *k-anonymity* was set to 5, refer to Section 4.1 for details on privacy metrics.) This is an iterative process in which micro-aggregation of continuous variables, recoding (e.g., super-categories or recoding sparsely populated extremes into bands), local suppression, post-randomization, adding noise or shuffling methods are applied until the pre-specified threshold for k-anonymity and l-diversity is reached.

In our experiments in Section 6 we do not use the actual anonymised dataset released by CCHIC. We instead generate an anonymised dataset that emulates it using the pipeline described in Stadler et al. [2021] (constraining k-anonymity to 5, among other things).

3 Synthetic data generation

Generating a synthetic dataset typically involves fitting a generative model to the original (sensitive) data, and then sampling from this model to produce synthetic records. There different types of models but in this work we focus on PrivBayes, a Bayesian synthetic method. We leave other methods (e.g. deep learning) as the topic of future work.

3.1 PrivBayes

PrivBayes [Zhang et al., 2014, 2017b] is a differentially private method to approximate the joint probability distribution of a sensitive data set. It uses a set of conditional probabilities between lower dimensional subsets of features from the original data set. The method constructs a Bayesian network connecting each feature of the original data set to up to k parent features (as a hyperparameter) and estimates the conditional probability distribution for each feature and its restricted set of parent features. Both the determination of the network structure and the estimation of the conditional probability distributions are done in a differentially private way. Synthetic data sets can then be generated by sampling from the differentially private approximate distribution represented by this Bayesian network.

In addition to the differential privacy budget ϵ (see Section 4.2), PrivBayes has two key parameters. The first is k, the maximum number of parents each feature is permitted in the structure of the network. The second is β, the proportion of the differential privacy budget that is spent on determining a good network structure versus estimating the conditional probability distributions. In this work, we use the PrivBayes implementation from Ping et al. [2017].

4 Privacy metrics

Here we introduce some common methods to quantify the privacy of individual-level datasets and justify our choice of using intruder attacks in this report. For a more thorough review of privacy measures for synthetic data and other applications, the reader is referred to Wagner and Eckhoff [2018].

4.1 K-anonymity, l-diversity and data categories

K-anonymity is one of the most widely used methods to guarantee and quantify privacy in datasets. It counts the number of individuals that sharing an intersection of key variables. For example, if within a dataset the sets of individuals with matching sex, race and ethnicity have a minimum size of $k = 10$, then k-anonymity is 10 for these attributes in the dataset. The higher this minimum value, the higher the privacy of individuals is protected.

The concept of k-anonymity only applies to categorical variables where we can group individuals. Therefore, in the CCHIC data, continuous key variables are converted to categorical ones by aggregating. Sizes of buckets for aggregation are initially small to minimize information loss, but they will be increased such that a pre-specified k-anonymity (and l-diversity) are reached, that is, coarsening the data to increase k-anonymity. However, changing the bucket size is not the only method to achieve a given k-anonymity. Other methods include micro-aggregation of continuous variables, recoding, local suppression, post-randomization, adding noise or shuffling methods.

L-diversity counts how varied other sensitive fields are within a k-anonymous group. A certain level of L-diversity protects from attribute inference based on the fact that some sensitive attributes might have the same value for all individuals in a group that shares the same key variables.

More details about how k-anonymity is used in the CCHIC dataset can be found in the CCHIC anonymisation SOP[13]. Based on these concepts, the data fields were classified into four categories: direct identifiers, key variables, sensitive fields and non-identifying variables.

4.2 Differential privacy

Differential privacy (DP) provides a robust, meaningful, and mathematically rigorous definition of privacy [Dwork et al., 2006]. Consider two databases, D_1 and D_2, that differ only on a single entry (e.g., where one patient is removed from CCHIC). These two databases are neighbors (aka parallel or adjacent). We denote a randomized algorithm by M which takes a dataset as input (i.e., the actions of the trusted party holding the data) and outputs value from some output space O. This randomized algorithm M is (ϵ, δ)-differentially private if for all $S \in O$ and for all neighboring datasets D_1 and D_2:

$$\mathbb{P}[M(D_1) \in S] \leq e^{\epsilon}.\mathbb{P}[M(D_2) \in S] + \delta \tag{1}$$

[13] https://github.com/CC-HIC/ccanonym/blob/master/inst/SOP%20data%20release.pdf

The probability $\mathbb{P}$ is taken with respect to the randomness used by the algorithm M. It should be noted that this equation should hold for all the neighboring datasets (i.e., that differ on a single element) and all subsets $S \in O$. Intuitively, for some query on a dataset D_1, a differentially private algorithm produces an output, regulated by ϵ and δ, that is statistically indistinguishable from the same query on the neighboring dataset D_2. In other words, the above equation states that inclusion or exclusion of a particular sample in the dataset changes the probability of an outcome by a multiplicative factor of e^{ϵ} and an additive amount of δ. ***PrivBayes***, described in Section 3.1, is one of the algorithms in the literature that satisfy this definition.

The value of ϵ constitutes a *bound* on privacy. It might be possible to determine a better bound for a given method, meaning that the method offers better privacy protection than indicated by ϵ. From the point of view of a user of data synthesis methods, who might have a particular 'privacy budget'—a maximum ϵ that they consider acceptable—the best published bound is the strongest privacy claim that they can make.

4.3 Intruder attacks

Despite the formal privacy guarantee offered by differential privacy and its increasing influence in synthetic data literature and practice, there challenges associated with its use. The interpretation of the ϵ bound is not straightforward; it is hard for non-expert users and the public to understand what this bound practically means for the level of privacy they should expect. In addition, the best choice of ϵ for a particular privacy application is not clear even for expert users [Stadler et al., 2021, Wagner and Eckhoff, 2018]. Ideally, we would like a way to translate a certain privacy bound to a more interpretable measure.

In addition, differential privacy cannot be applied as a privacy measure to the anonymised version of the CCHIC dataset (and in fact any dataset anonymised using traditional methods); it is a quantity that is defined as part of the synthesis process in advance, rather than a data-driven metric calculated from the release data.

On the other hand, k-anonymity and similar measures, despite being data-driven, cannot guarantee privacy in certain scenarios, i.e. even datasets with high values can be vulnerable to certain types of attack (e.g. homogeneity and background knowledge attack on k-anonymous datasets, skewness and similarity attack on l-diverse datasets [Rajendran et al., 2017].

In order to be able to compare synthetic and anonymisation methods, communicate our privacy assessment to non-experts and the public and avoid the caveats of k-anonymity and l-diversity, we need a privacy measure that is data-driven, easily interpretable and captures many types of risks. In this report, we quantify privacy by running two types of motivated intruder attacks against synthetic and anonymous datasets and calculating probabilities of success of these attacks and other related metrics. Each attack assumes that the intruder has specific prior knowledge. The intruder attack privacy metrics are purely data-driven, they are easily interpretable and they can be tuned in various ways to capture different intruder scenarios and assumptions and the associated risks.

In more detail, the attacks we use are the following:

4.3.1 Attribute inference attack

This is attack aims to infer a sensitive attribute value for a targeted individual or a set of individuals, e.g. their HIV or Cirrhosis status. We assume that the intruder has partial knowledge about the targeted individuals, i.e. they know all the other variables for these individuals (age, ethnicity, weight, height, admission date, etc). They also have access to the released dataset (either synthetic or anonymised) and they know the percentage of people in the ICU patient population who are HIV-positive or Cirrhosis-positive (prior knowledge).

This attack could simulate a scenario where a neighbour of a patient who might know their neighbour's age, ethnicity, etc and when they were admitted to hospital and they might try to infer sensitive health information. Or a generic actor who tries to do the same for a VIP (e.g. public figure, politician, celebrity).

The intruder's strategy is described in detail in Stadler et al. [2021]. The intruder has access to a partial record and the released dataset. If the released dataset is an anonymised dataset, they first try to match the partial record with records in the released data. If the match results in a unique record, they use this record's sensitive attribute as their guess, which has probability $p = 1$ of being successful. If there is no unique match or if the released data are synthetic, the intruder trains a supervised ML model to predict the sensitive value based on the other variables in the data. The released dataset is used for the training and then the partial target record is fed into the training model to make a guess.

4.3.2 Membership inference attack

This attack aims to infer if a targeted individual is a member of the original dataset used for generating the released (synthetic or anonymised) data. It assumes the intruder has access to the complete record (i.e. all variables) of the targeted individual. It makes the additional assumption that the intruder has access to a complete sample from the original dataset. Both these assumptions are possible but very strong; a real-world intruder might not have access to an original sample and might only know a patient record partially. Finally, the intruder has access to the released data and they know the synthetic method used to generate them (e.g. PrivBayes). We further assume that the data owner enforces a 50% chance of including each patient record in the synthetic training sample and that the intruder knows that.

The attack simulates a scenario where some actor has a data about an individual and they are trying to infer if a person was treated at one of the ICUs connected to CCHIC.

The intruder's strategy is described in detail in Stadler et al. [2021]. The intruder performs a so-called shadow model attack: They first sample a training data set from the sample of the original dataset they have access to. They train a synthetic model (same type as the one used by the data owner) using this training dataset and generate a set of synthetic datasets from it and assign them the label 0. They then repeat the process but this time they add the target record to the training

dataset. This results in another set of synthetic datasets to which they assign the label 1. Finally, they train a classifier (the *discriminator*) on the labelled synthetic datasets. The classifier takes as input a set of features from a dataset and guesses whether the dataset's label is 0 or 1, i.e. whether the target record was included in the training dataset. This classifier is then fed the released dataset and guesses membership for the targeted individual. Feature creation can be done in different ways which are described in Stadler et al. [2021].

5 Utility metrics

When releasing data to be used by the research community, these data should be useful for analysis at least up to a certain level. It would be pointless to release a dataset that is so distorted compared to the original data that does not allow any useful insights to be extracted from it.

Utility metrics capture this usefulness in different ways. A first category of metrics captures the overall distributional similarity between original and synthetic/anonymised data using marginal/joint distribution similarity measures like descriptive statistics, distance metrics, correlations, etc. These metrics are built to reflect the usefulness of the data in a broad sense, i.e. for any potential task that the user might want to perform. A second category captures similarity when performing specific types of analyses with the data and can be particularly useful when the exact tasks to be performed are known. A typical example is training ML models using the data and measuring predictive performance. Finally there are some attempts to combine these two types of metrics [Taub et al., 2020].

Here we use both types of metrics, specifically:

- Descriptive statistics: We calculate means and median for continuous variables and category frequencies for categorical variables and compare them between different synthetic, anonymised and original data.
- Predictive performace on ML tasks: We train random forest classifiers that predict four of the variables in the dataset ("Dead or alive on discharge", "Ethnicity", "HIV", "Cirrhosis"). For each model we use all other variables in the dataset as features. The model that predicts "Dead or alive on discharge" from all other variables, although simplistic and with limited predictive power, is representative of the types of tasks a clinical researcher might want to perform using a released dataset. Predictive performance is evaluated on a hold-out (test) sample from original data.

6 Experiments

This section described the experiments we performed to measure privacy and utility and the software used for the implementation.

6.1 Data preprocessing

The following preprocessing steps are applied to the original data before performing the experiments:

- Incomplete patient episodes (i.e. when the patient was still in the ICU at the time of data extraction) are dropped.
- Any "unk" (i.e. unknown) values in continuous variables are replaced with NaN (Not a Number).
- Any "unk" (i.e. unknown) values in the HIV and Cirrhosis variables are converted to "0.0" (i.e. negative).
- For any patients with height (in cm) under 2.80, the height is multiplied by 100 to convert to meters (these are cases where height was originally inputted in meters). After this conversion, any patients with height lower than 100 cm are removed.
- Any patients with weight lower than 30 kg are removed.

6.1.1 Target records

Seventeen individuals, with different characteristics, were selected for the experiments as listed in Table 1. Here, we define "outlier" as those with significantly different weight, height, age or ethnicity from others.

Table 1: Eleven individual patients included in the experiments as special targets. They are selected based on weight, height, age, ethnicity, HIV, Cirrhosis or survival status. Some of them are used in both attribute and membership inference attacks as indicated in the last column.

IDs	**Description**	**Used in**
A, H, J	Average record	Both attacks
B, I, K	Outlier record	Both attacks
F	Average record, HIV+	Both attacks
E	Outlier record, HIV+	Attribute inference
C	Average record, Cirrhosis+	Both attacks
G	Outlier record, Cirrhosis+, Died	Attribute inference
D	Average record, HIV+, Cirrhosis+	Attribute inference

6.1.2 Generative mechanisms used in experiments

For both intruder attacks, we ran experiments separately for the following data release (generative) mechanisms:

- **PrivBayes:** We used the implementation contained in Stadler et al. [2021] and developed a fork of the authors' repository in GitHub[14] which contains several modifications to adapt the pipeline to the experiments in this report. We separately ran experiments for $\epsilon = 0.01$,

[14] `https://github.com/alan-turing-institute/synthetic_data_release`

$\epsilon = 0.1$, $\epsilon = 10.0$ and $\epsilon = 30.0$. We set hyperparameters so that each feature has at most two parent features in the Bayesian network ($k = 2$) and the maximum number of categories for categorical features is 25.

- **Anonymisation:** We used the anonymisation (or sanitization) algorithm described and implemented in Stadler et al. [2021]. This is modeled after the methodology described in NHS England's A&E synthetic data publication [nhs]. We chose k-anonymity to be 5 for variables "sex", "ethnicity" and admission date variables to emulate the anonymisation process used by CCHIC for their publicly released dataset. We also set the number of buckets for categorical features to 10 (used for aggregation/grouping) and capped numerical values to each variable's 99% quantile.

In addition to these generative mechanisms, we use the following additional baseline guessing methods which are based on prior knowledge and do not make use of released data: For the attribute inference attack, we assume the intruder knows the percentage of HIV- and Cirrhosis-positive patients in ICU and they just guess whether each target is positive based on this prior. The prior percentages are calculated by averaging the "HIV" and "Cirrhosis" variables in the original data within CCHIC. For the membership attack, we assume the intruder can make a random guess about the membership of each target as they know there is a 50% probability of membership for each target.

6.1.3 Detailed description of experiments

In order to capture the likelihood of successful attacks on data released from CCHIC, we run experiments where the various sources of randomness in the process of data generation/release and the process of intruder attack are captured and reflected on success metrics by running independent iterations with different random seeds.

The sources of randomness in data generation include the selection of the training sample from the original data (if a subset is used), the synthetic method training process and the generation of samples from the trained synthetic model. The sources of randomness in the intruder attacks include the training of the attribute inference model (for attribute inference attacks) and the training of the discriminator (for membership inference attacks).

Attribute inference In the attribute inference experiments, we separately ran the attack for the sensitive attributes "HIV" and "Cirrhosis". We targeted the complete set of records in the original database but also present results for specific records listed in table 1 which represent positive/negative cases, average individuals and outliers.

We run $n = 10$ iterations of the experiment, which simulates both the actions of the data owners when releasing the data and the actions of the intruder. Each iteration involves the following steps:

- Randomly select a subset of the original data. Subsets contain $l = 20,000$ individuals' records each.
- Train a synthetic model using the subset.

- Generate $m = 5$ synthetic datasets from the trained model (size $l = 20,0000$ each).
- (Intruder attack) For each synthetic dataset, train a Random Forest classifier to guess the sensitive values from the other variables in the dataset. We use the classifier as implemented in the Scikit-learn library [Pedregosa et al., 2011]. After training, guess the sensitive value for each target and for the whole original dataset subset. The guess involves taking the likelihood of each category as produced by the random forest classifier, multiplying it with the prior knowledge about the prevalence of HIV/Cirrhosis in the ICU population to get the posterior distribution. We then choose the category with the highest posterior probability.
- Evaluate success of the set of intruder guesses for each target separately using the set of guesses from the m synthetic datasets and calculating Accuracy, Balanced Accuracy, Precision, Recall and F1 metrics. Also evaluate the same metrics when guessing all targets in the subset of the original dataset of size $l = 20,000$.

After all n iterations finish, we calculate the mean and standard deviation of all success metrics and present them in Section 7. We follow the same process for anonymised datasets with the only difference being the generative mechanism for data release

Membership inference In the membership inference experiments, we separately ran the attack for each target record listed in table 1 which represent average individuals and outliers.

For *each targeted record* available to the intruder, the intruder first trains discriminators as follows:

- Randomly select a shadow sample from the original data (assumed to be available for the intruder). The size of this sample was set to $r = 11,000$.
- Randomly select $t = 5$ samples from the shadow sample, each of size $l = 10,000$.
- For each of the t training samples, fit the PrivBayes generative model without including the target record in the sample. For each of the t generative models, generate $m = 5$ synthetic datasets (size $l = 10,0000$ each) and attach the label 0 to them.
- Repeat the previous step but this time including the target record in the training sample and attaching the label 1 to the synthetic output datasets.
- For all synthetic datasets, extract feature sets from them using three independent feature algorithms: Naive, Correlations and Histogram (described in Stadler et al. [2021]). Train the discriminator classifier (a Scikit-Learn random forest) using the features and the labels assigned previously (independently for each feature set).

After the discriminators have been trained for all targets and feature algorithms, we run $n = 10$ iterations of the experiment for *each target*, which simulate the actions of the data owner when releasing the data and the actions of the intruder. For each target and iteration the following steps are performed:

- Randomly select a sample from the original data. The size of this sample was set to $u = 10,000$.

- Train two synthetic models using the subset, one without the target and one including the target.
- Generate $m = 5$ synthetic datasets from each trained model (size $l = 10,0000$ each).
- (Intruder attack) For each synthetic dataset, run each of the feature algorithms. Feed each resulting feature set to the respective discriminator trained previously to generate a guess about membership of the target being attacked.
- Evaluate success of the set of intruder guesses for this target using the set of guesses from the $2m$ synthetic datasets and calculating Accuracy, Precision, Recall and F1 metrics. This is done separately for each feature algorithm.

After all n iterations finish, we calculate the mean and standard deviation of all success metrics and present them in Section 7. We follow the same process for anonymised datasets with the only difference being the generative mechanism for data release.

7 Results

7.1 Attribute inference

We first examine attribute inference attacks. Figures 1 and 2 show attack success metrics calculated when guessing the HIV and Cirrhosis status of all the targets in the original data ($l = 20,000$ records) by looking at the released data. The metrics used are standard metrics in ML literature and can be understood even by non-experts, particularly if given simple examples (similarly to what was done in the CCHIC SOP report[15]). All generative mechanisms are shown (with PrivBayes ϵ ranging from 0.01 to 30.0).

The left plot in Figure 1 shows the accuracy, i.e. percentage of guesses (either positive or negative) that were correct. The accuracy is already very high when using only prior information and no data (~ 0.982). This is because the dataset is very imbalanced with very few HIV positive cases; it is easy to get a high accuracy if predicting positive almost always. With PrivBayes synthetic data as well as with anonymised (sanitised) data, the accuracy is slightly higher (~ 0.99) except when using $epsilon = 0.01$ and $\epsilon = 0.1$ in PrivBayes. High accuracy implies that the released datasets somewhat recreate the imbalance in the original data and therefore allow the intruder to train a model that predicts positive most of the time. In the low-epsilon PrivBayes releases, accuracy is lower with a large variance between independent runs, indicating that these synthetic datasets distort the dataset so much that they confuse the intruder; accuracy is lower even compared to not having any data at all (prior guess) and it seems to get closer to a random guess as ϵ bound decreases. The pattern of increasing privacy when ϵ decreases is the expected one for differentially private synthetic algorithms.

Nevertheless, accuracy gives us a limited understand of the success of the guessing task. It combines positive and negative HIV cases into one metric and does not give visibility on how many of the

[15] `https://github.com/CC-HIC/ccanonym/blob/master/inst/SOP%20data%20release.pdf`

positive cases were guessed correctly (i.e. recall or true positive rate) or how many of the positive guesses were successful (i.e. precision or positive predictive value). Looking at the middle plot in Figure 1 it is clear that, despite the high accuracy, precision is close to zero (mean ranging from 0 to 0.022 with low variance) for all synthetic algorithms. This means that only a small minority of the intruder's positive guesses are successful. For comparison, with prior-only information, 7.7 out of $1,000$ positive guesses are actually HIV-positive and with PrivBayes this ranges from 6.3 to 22.5 among $epsilon$ bounds (mean values with standard deviation of $6.7 - 45.2$). Practically, for this type of attack the intruder is not be able to have confidence in their guesses and harm to patients is unlikely.

The same plot shows the recall metric which is also close to zero for prior guess and the two PrivBayes instantiations with the highest ϵ bounds. For low ϵ bounds recall increases because the synthetic data are distorted to the extent that HIV status is no longer imbalanced; this in turn makes the intruder's ML model predict a lot of positives. Some of these positives are true positives and this is why recall increases (a larger percentage of the true positives are captured). But this is of little use to the intruder as precision is still low and therefore they cannot tell which guesses are correctly labelled positive.

Finally, we observe that the precision of the anonymised release is 1.0 which means that all positive predictions are actual positives. The recall is 0.166. This means that more than 16% of the HIV-positive cases are guessed correctly and these are the only positive guesses, therefore the intruder gains a significant advantage. This shows that the anonymisation algorithm used in these experiments is not protecting individual privacy adequately.

The right plot in Figure 1 shows F1 metrics for the same data. The F1 score is the harmonic mean of precision and recall and is designed to be sensitive to either one of them having a small value. The binary F1 score calculates the metric treating HIV-positive as the positive label. We observe that its value is close to zero and close to the prior guess score (as expected based on precision and recall scores), with a very slight increase for low epsilons due to the higher recall score. The anonymisation F1 score is clearly higher. We also show the macro-averaged F1 score which calculates F1 for both HIV-positive and HIV-negative and takes the unweighted average. This partially captures the effect of true negatives which are not considered with the other metrics; F1-macro decreases with lower ϵ bounds due to the high number of HIV-negative cases labelled positive. Finally F1-micro is identical to the accuracy metric.

Results for the inference attack against the "Cirrhosis" sensitive attribute produce very similar results shown in Figure 2.

Finally, it is worth noting that the batch attack described above is an unlikely event; the intruder would have to have access to the partial records of all $l = 20,000$ patients to be able to benefit from the average marginal gains described above. In practice the intruder is more likely to have access to one or a few partial records.

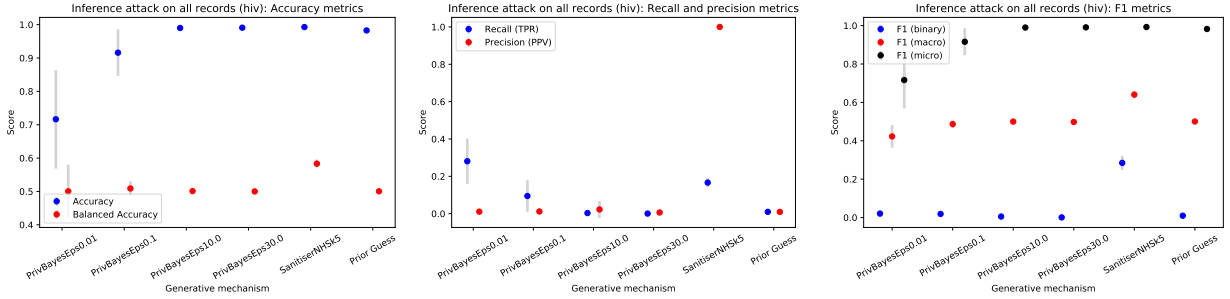

Figure 1: Accuracy, Balanced Accuracy, Recall, Precision and F1 metrics (binary, macro, micro) achieved by intruder when inferring the HIV attribute for all the records in the training data ($l = 20,000$). Metrics are plotted for all generative mechanisms used for data release and for prior guesses made without data (based on prior knowledge about HIV prevalence in ICU population). The grey bars represent +1/-1 standard deviation. PrivBayes ϵ values used are $[0.01, 0.1, 10.0, 30.0]$.

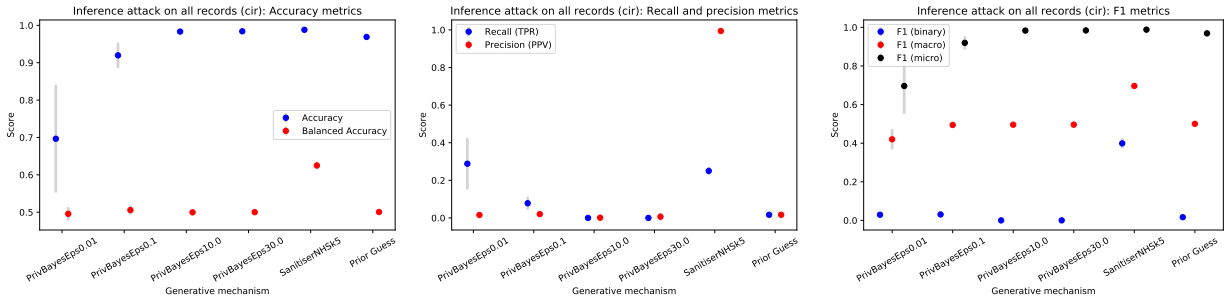

Figure 2: Accuracy, Balanced Accuracy, Recall, Precision and F1 metrics (binary, macro, micro) achieved by intruder when inferring the Cirrhosis attribute for all the records in the training data ($l = 20,000$). Metrics are plotted for all generative mechanisms used for data release and for prior guesses made without data (based on prior knowledge about Cirrhosis prevalence in ICU population). The grey bars represent +1/-1 standard deviation. PrivBayes ϵ values used are $[0.01, 0.1, 10.0, 30.0]$.

In addition to the above results, we also present Figures 3-6 which show the accuracy achieved for some chosen patients across independent runs. These figures show the accuracy the intruder is expected to have in predicting the sensitive attribute for a particular target if they ran the attack many times on data releases made with different random seeds. Figure 3 shows accuracy for HIV-positive patients idD, idE and idF. Patients idD and idF are typical/average records, i.e. close to the average in terms of age, weight, height and ethnicity. idE is an outlier in terms of ethnicity and weight and has significantly below-average age. Mean accuracy is zero in all the synthetic experiments. This is in line with the previous figures, showing that it is really hard for an intruder to successfully guess HIV-positive cases, even when attacking outliers. The only release mechanism where intruder accuracy is higher for one of the targets is the anonymisation mechanism which has mean accuracy 0.35, which is also in line with previous figures. Prior guesses have low accuracy as they are rarely successful by chance. Figure 4 shows the same metric for CIR-positive

targets idC, idD (average patient) and idG (outlier). Again, accuracy is zero except for PrivBayes ($\epsilon = 0.01$) where it is very low (~ 0.01 with small standard deviation). The same patterns are observed for other mechanisms.

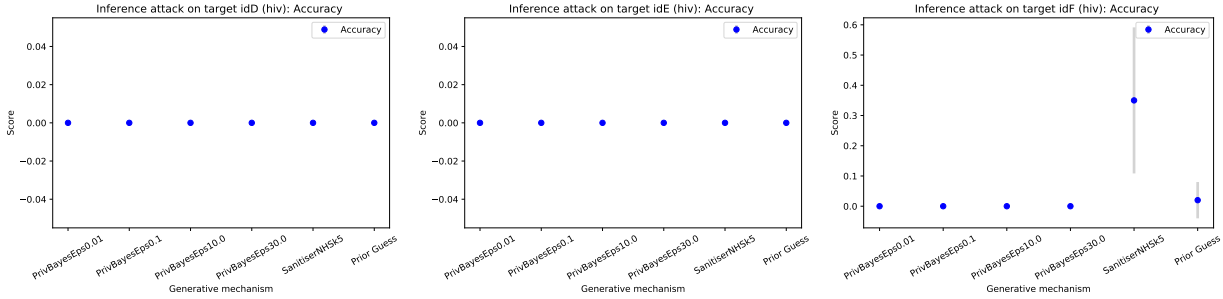

Figure 3: Accuracy achieved by intruder when inferring the HIV attribute for HIV-positive records idD, idE and idF. idD and idF are typical records, i.e. close to the average in terms of age, weight, height, ethnicity. idE is an outlier in terms of ethnicity and weight and has significantly below-average age. Accuracy is plotted for all generative mechanisms used for data release and for prior guesses made without data (based on prior knowledge about HIV prevalence in ICU population). The grey bars represent +1/-1 standard deviation. PrivBayes ϵ values used are $[0.01, 0.1, 10.0, 30.0]$.

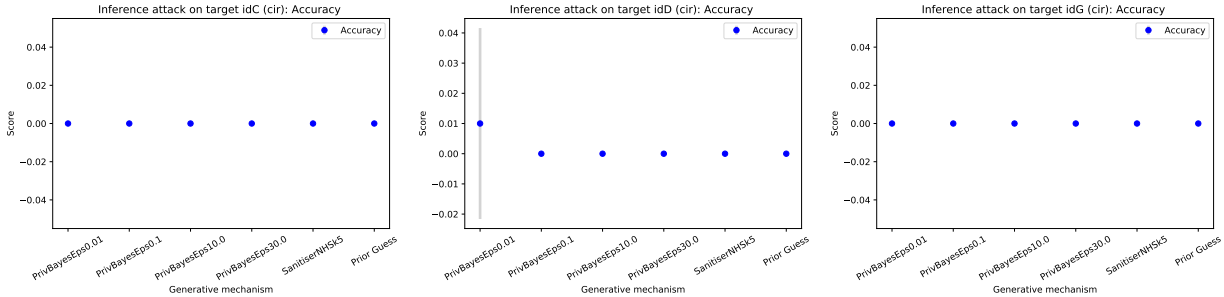

Figure 4: Accuracy achieved by intruder when inferring the Cirrhosis attribute for Cirrhosis-positive records idC, idD and idG. idC and idD are typical records, i.e. close to the average in terms of age, weight, height, ethnicity. idG is an outlier in terms of ethnicity, height, weight, has significantly below-average age and was deceased on discharge. Accuracy is plotted for all generative mechanisms used for data release and for prior guesses made without data (based on prior knowledge about HIV prevalence in ICU population). The grey bars represent +1/-1 standard deviation. PrivBayes ϵ values used are $[0.01, 0.1, 10.0, 30.0]$.

Figures 5 and 6 show accuracy for HIV-negative and Cirrhosis-negative patients (both average cases and outliers). This is almost always 1.0 for synthetic methods as the intruder's model guesses negative most of the time (with the exception of PrivBayes $\epsilon = 0.1$ for target idG). Prior guesses also score close to 1.0 as without data the intruder guesses negative most of the time based on the prior information.

Summary The results demonstrate that a motivated intruder trying to infer HIV or Cirrhosis status from synthetic data generated from this CCHIC subset would perform very marginally better than based on prior information. Their success rate when predicting positive cases (i.e. precision) would be low for all PrivBayes configurations (mean of 1.2-22.5 successes per $1,000$ positive guesses with a standard deviation of $1.8 - 45.2$). It is unlikely that they would be able to infer positive HIV/Cirrhosis status with confidence. This applies to average cases and outliers. For differentially private algorithms with low ϵ bound, the intruder would also fail to predict negative cases to some extent. In contrast, the anonymisation method used in these experiments (which has differences compared to the one used in CCHIC) was not able to protect the sensitive attributes effectively, leading to many successful positive cases being guessed.

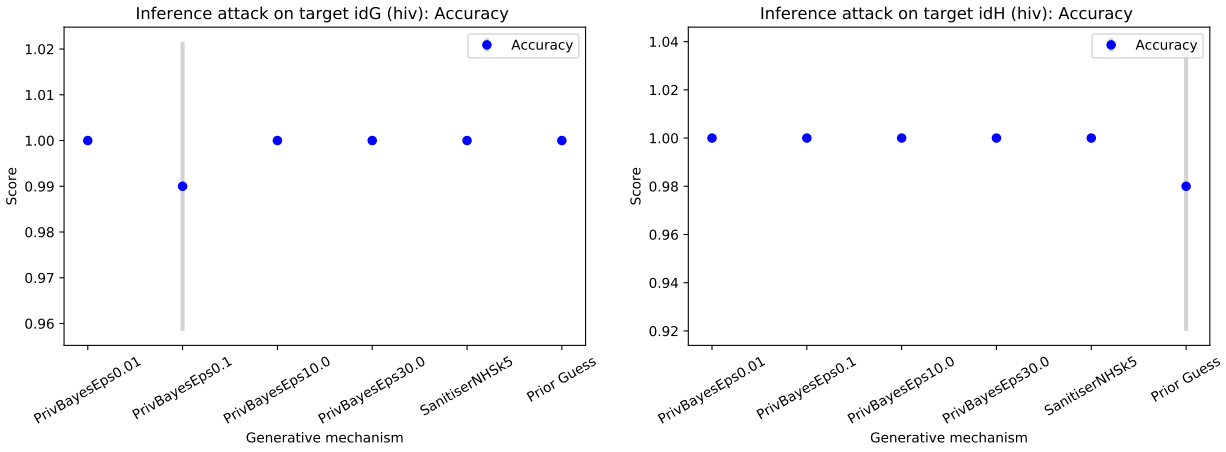

Figure 5: Accuracy achieved by intruder when inferring the HIV attribute for HIV-negative records idG and idH. idH is a typical record, i.e. close to the average in terms of age, weight, height, ethnicity. idG is an outlier in terms of ethnicity, height, weight, has significantly below-average age and was deceased on discharge. Accuracy is plotted for all generative mechanisms used for data release and for prior guesses made without data (based on prior knowledge about HIV prevalence in ICU population). The grey bars represent +1/-1 standard deviation. PrivBayes ϵ values used are $[0.01, 0.1, 10.0, 30.0]$.

7.2 Membership inference

We next examine the outcome of the membership inference attack experiments. In these experiments, the intruder trains their distinguisher multiple times and attacks multiple randomly generated synthetic datasets, half of which contain the target record and half do not. Figures 7 and 8 show the accuracy score and precision score averaged across all targets in Table 1 as well as the same scores for two individual targets (the average case idF and the outlier idI). Note that the averaged scores are not representative of the whole dataset but only of the set of targets selected here; experiments covering the whole dataset were not performed due to limited computational capacity within the DSH environment.

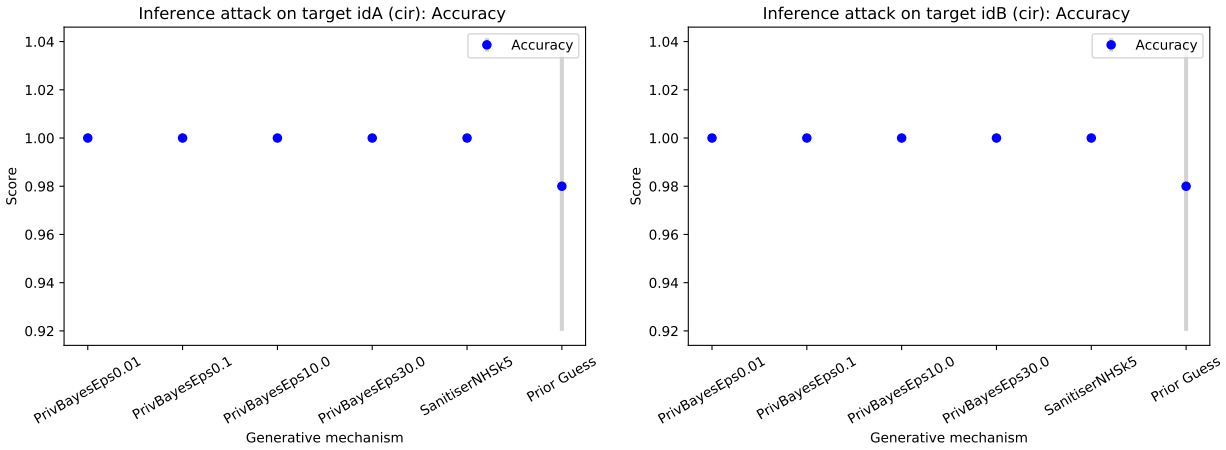

Figure 6: Accuracy achieved by intruder when inferring the Cirrhosis attribute for Cirrhosis-negative records idA and idB. idA is a typical record, i.e. close to the average in terms of age, weight, height, ethnicity. idB is an outlier in terms of age and ethnicity. Accuracy is plotted for all generative mechanisms used for data release and for prior guesses made without data (based on prior knowledge about HIV prevalence in ICU population). The grey bars represent +1/-1 standard deviation. PrivBayes ϵ values used are $[0.01, 0.1, 10.0, 30.0]$.

The averaged plot (Figure 7) indicates that, on average, the attack achieves higher F1 score than a random guess (which would score 0.5). This applies both to the synthetic and anonymised data and to all feature set methods. There is no consistent change in F1 when changing the ϵ bound which seems to indicate that the amount of noise inserted by differential privacy makes little difference to the protection against membership inference attacks. This is a similar pattern to the one observed for attribute inference attacks. Overall, F1 ranges from 0.598 to 0.685 for PrivBayes, with standard deviations in $0.221 - 0.247$. The anonymised dataset has an average F1 of 0.696 with standard deviation 0.095, therefore providing similar protection to synthetic data but with lower variance across runs and targets.

Average precision, which tells us how many of the positive guesses (i.e. guessing that the target is a member of the dataset) are actual positives, takes values around 0.5 (and more often marginally above 0.5) in all scenarios. This indicates that the intruder's positive guesses have similar success to a random guess and is probably a more interpretable and useful metric than F1. Specifically, the average precision ranges from 0.459 to 0.538 for PrivBayes (with 3 out of 4 values higher than 0.5) and the standard deviation range is $0.351 - 0.378$. For the anonymised data, the average precision is 0.544 with standard deviation 0.142, close to the synthetic data but with lower variance. The above results indicate that if the intruder had complete records for a large number of targets, 50 out of 100 of their positive membership guesses would be successful with a random guess but this would increase to up to 53 on average (or might even drop slightly) with the synthetic/anonymised data. This is a small improvement but most importantly the variance is so big that they would be very

uncertain about this success rate; depending on the particular random seeds used to train their attack and the seeds used by data owners to generate data, the rate could change dramatically.

Additionally, intruders are unlikely to be able to access a large number of complete records and benefit from the marginal average gains demonstrated here (i.e. guess membership correctly for an increased percentage of records which might mean several more correct guesses). The more likely scenario is of an intruder that has access to a few records and attacks those only, in which case it is not clear how they would be able to benefit in practice given the uncertainty around the guesses. Finally, even the assumption of a complete record being available to the intruder can be unrealistic in many scenarios, e.g. knowing the date of admission but not knowing if the patient is a member of the dataset is possible (e.g. if they do not know the hospital they were admitted to) but this scenario might bring limited benefits to the intruder.

Summary The results demonstrate that a motivated intruder trying to infer membership of individual patients from synthetic data generated from this CCHIC subset would have marginally higher rate of correct positive guesses over all positive guesses compared to random guess (maximum 0.53 vs. 0.50) and an F1 score which is higher than random guess on average (up to 0.68 vs 0.50). The anonymised dataset has similar privacy protection. All scores have large variance; the intruder sometimes achieves very high F1 and precision (even 1.0) while in other cases they score 0.0. While released data might bring marginal gains for an intruder performing a batch attack on many records, they are unlikely to help a lot in the more realistic scenario of targeting a few leaked records, as the intruder would have very low confidence in their guesses due to variance. Nevertheless, the difference in F1 scores between synthetic methods and random guess show that the intruder gains *some* useful information from the synthetic data when doing a membership inference attack.

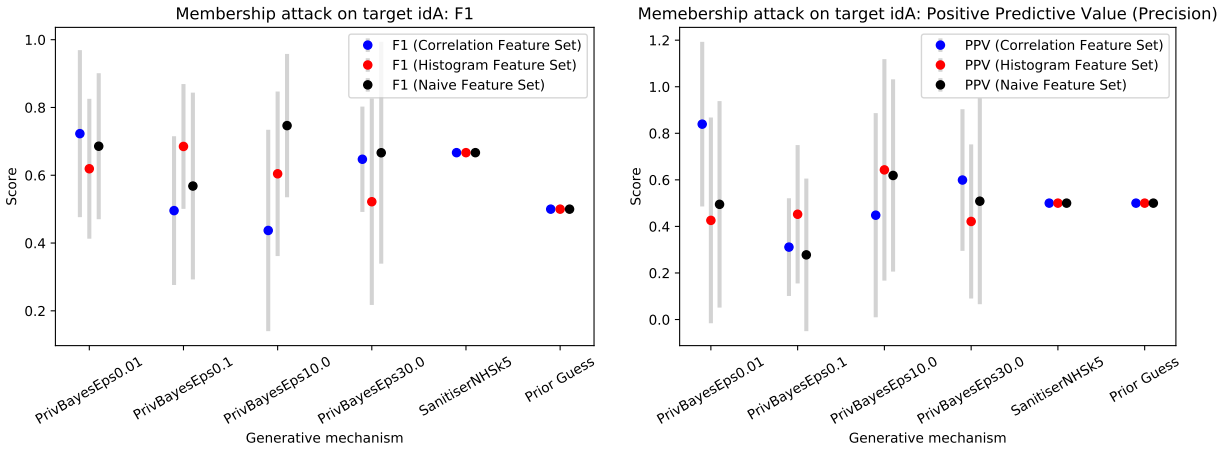

Figure 7: Average F1 score and precision achieved by the intruder when inferring membership for all targets in Table 1. Metrics are plotted for all generative mechanisms used for data release and for prior guesses made without data (based on random guess). The grey bars represent +1/-1 standard deviation. PrivBayes ϵ values used are $[0.01, 0.1, 10.0, 30.0]$.

Individual targets' metrics (Figure 8) differ substantially in some cases, e.g. precision or accuracy can shoot up or down for particular ϵ values or feature set methods. But there is no clear pattern when comparing averages vs. outliers or across different feature methods. And the variance in metrics is large even for individual patients. Similar results are observed for other individual cases from Table 1.

Summary This is due to the more deterministic nature of the anonymisation process versus the synthetic process.

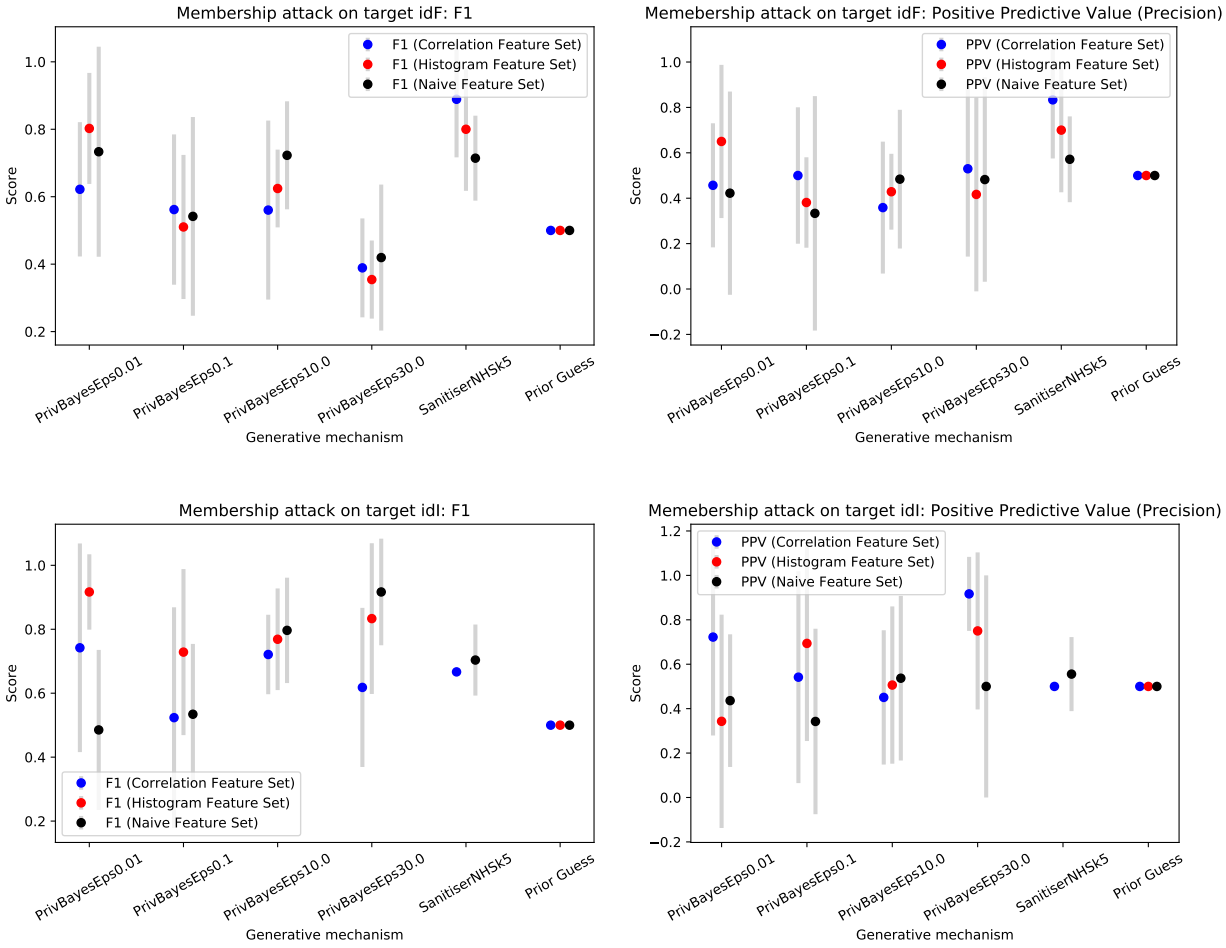

Figure 8: F1/precision when inferring membership for targets idF (average target - top row) and idI (outlier - bottom row). Metrics are plotted for all generative mechanisms used for data release and for prior guesses made without data (based on random guess). The grey bars represent +1/-1 standard deviation. PrivBayes ϵ values used are $[0.01, 0.1, 10.0, 30.0]$.

7.3 Utility

Figure 9 shows the out-of-sample predictive performance of four classifiers trained on the datasets produced by all synthetic and anonymisation algorithms and on the raw data. This is a standard

way to measure utility in synthetic data literature, as a common task researchers want to do with released data is to train a predictive model. A possible use of this dataset could be predicting the "das" variable from all other variables. The results show that the accuracy and macro-F1 score of the PrivBayes datasets increase with ϵ as expected and their values are close to the values of the raw data for larger ϵ bounds. There is some variance for lower ϵ bounds. The predictive performance is similarly high for the anonymised dataset. The same pattern is observed when predicting other variables including the binary variables "hiv" and "cir". For the latter two it is clear that we observe similar results to what we saw in the attribute inference attack section as we are effectively performing the same task. In general, it seems that synthetic data have decent utility that would allow some ML models to be trained with similar results to the raw data when the ϵ bound is 10.0 or 30.0 (note that for the "eth" variable accuracy with synthetic data is lower even for $\epsilon = 10.0$). For lower ϵ bounds the distortion in the dataset affects accuracy and F1 noticeably.

Figure 10 shows the means of numerical variables in the dataset, which is another way to capture utility of the dataset. The plots demonstrate how higher ϵ values lead to better (closer to raw) estimates of the mean of all variables. Means also have less variance as the ϵ bound increases. There is a substantial deviation from the correct mean of "wgt" with all synthetic datasets. The other two variables are affected less. The anonymised dataset seems to capture "wgt" much better but deviates more in the other variables. It is worth noting that the anonymised dataset's mean values have very small variance while the synthetic ones vary a lot between independent runs. This variance makes the utility of the synthetic data more uncertain for the user compared to anonymised data. The results when measuring the medians of the same variables are similar and are not shown here.

Finally, Figure 11 shows the frequencies of four categorical variables as measured from all released datasets and the raw dataset. The "hiv" plot shows how much low ϵ bounds distort the distribution of this variable, making it almost 50% positive for the lower ϵ. This is the reason the attribute inference attack guesses a lot of positive values for low ϵ, leading to a relatively high recall score. The frequencies converge to the correct ones as ϵ increases and the anonymised versions is also close to the raw frequencies. Similar trends are observed for the "das" variable.

For the "eth" and "wad" variables which have more categories we can see that the highest two ϵ bounds again come closest to capturing the raw distribution but lower ϵ bounds and the anonymised version do not. In fact, the $\epsilon = 0.01$ and the anonymised version are excluded from the "eth" plot because they do not manage to generate predictions for all the categories. Frequency variance for low ϵ bounds is large for all variables.

Summary The utility experiments show that the synthetic data maintain some utility (both task-focused and generic distribution-focused) when the differential privacy bound is not too low. It seems that they can be used for at least some research tasks and substitute raw data for parts of analyses. Nevertheless, they capture some variable distributions better than others and performance varies substantially with the random seed used. Anonymised data seem to also perform well in respect of utility and with less variance than synthetic data but distort variables with multiple classes more than synthetic data do. Having said that, the variable subset and the ML tasks defined here are

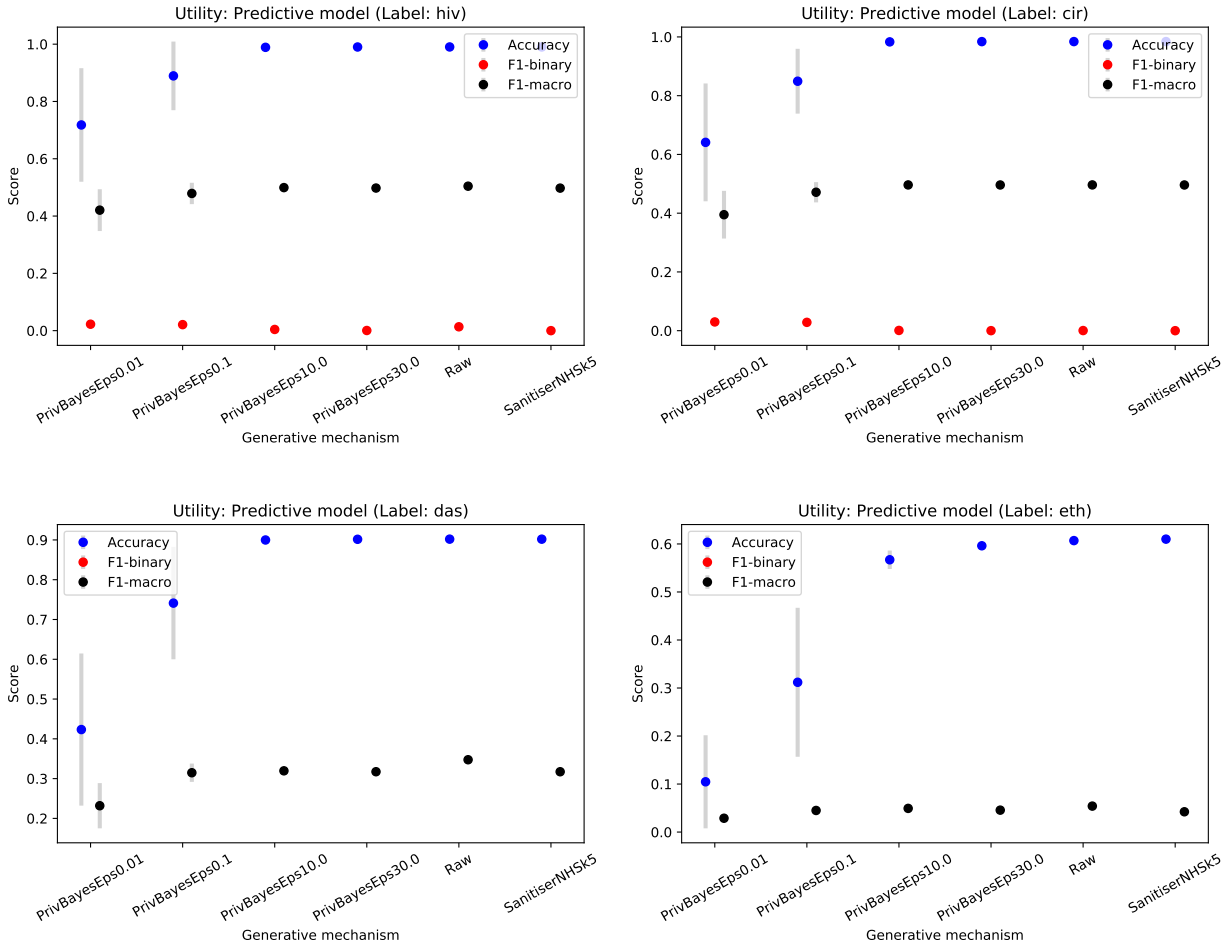

Figure 9: Utility: Out-of-sample predictive performance of ML classifiers trained on released data. Classifiers predict one variable from all other variables in the dataset. HIV and Cirrhosis are binary, "das" and "eth" are multi-class. Accuracy, F1 (binary) and F1 (macro) metrics are plotted for all generative mechanisms used for data release and for the raw data. F1 (binary) is shown only for binary variables. The grey bars represent +1/-1 standard deviation. PrivBayes ϵ values used are $[0.01, 0.1, 10.0, 30.0]$.

not as complex as a real research scenario and do not include many of the time-series variables in the CCHIC dataset that would be most interesting for researchers to work with.

8 Conclusions

Combining the results from the privacy and utility experiments of the previous section, it seems that CCHIC synthetic data generated with the PrivBayes algorithm (with ϵ above a certain threshold value which might be 0.1 here) can substitute raw data to some extent when performing research tasks. The outcomes of specific ML tasks using these datasets but also some of their distributional characteristics are not very far from those of raw data. Nevertheless, there are cases where some

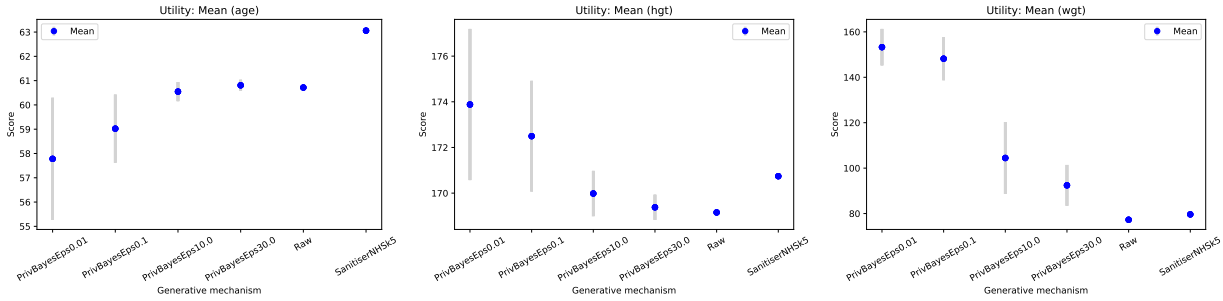

Figure 10: Utility: Variable means of numerical variables "age", "height" and "weight" for all generative mechanisms used for data release and for the raw data. The grey bars represent +1/-1 standard deviation. PrivBayes ϵ values used are $[0.01, 0.1, 10.0, 30.0]$.

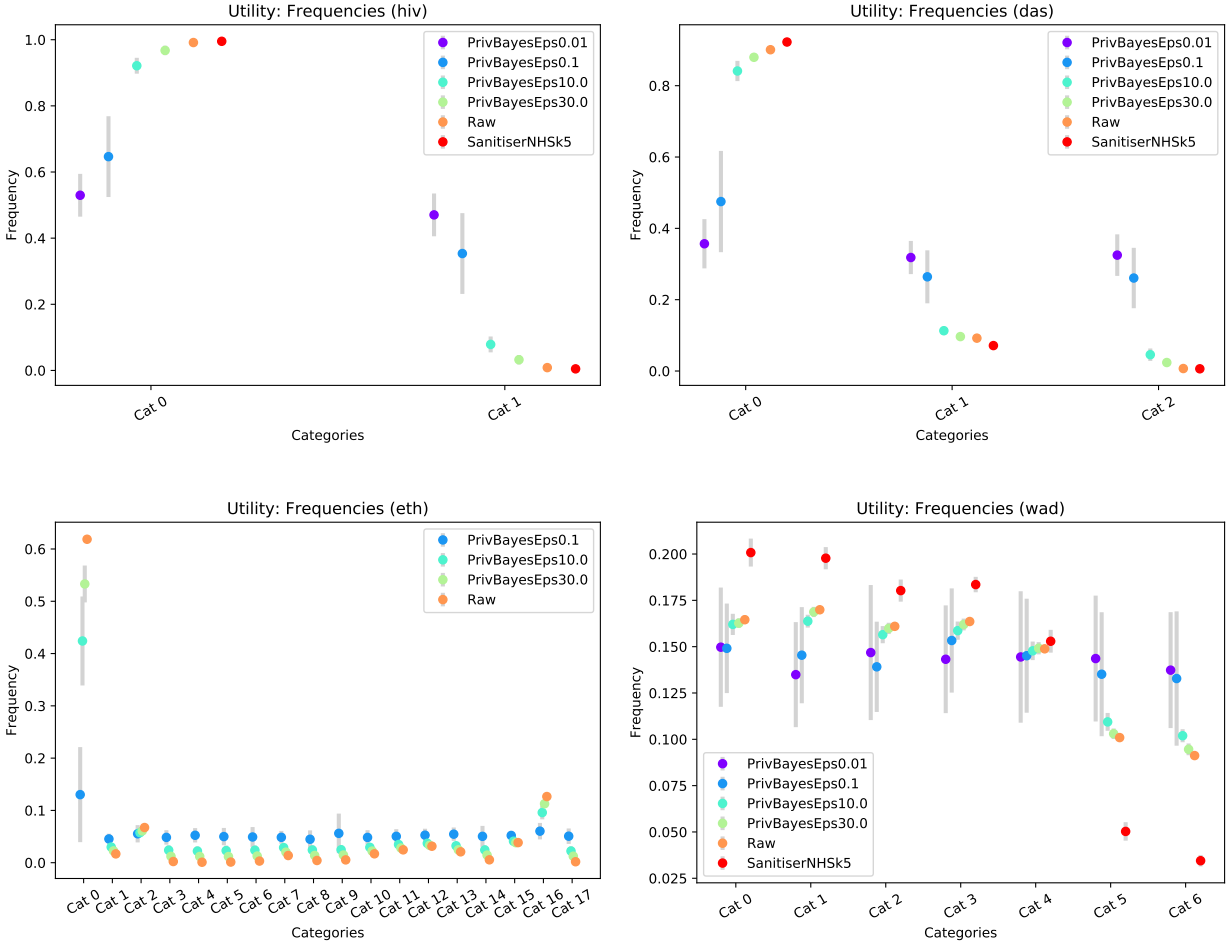

Figure 11: Utility: Categories' frequencies of categorical variables "hiv", "das", "eth" and "wad" for all generative mechanisms used for data release and for the raw data. Categorices are integer indexed. The grey bars represent +1/-1 standard deviation. PrivBayes ϵ values used are $[0.01, 0.1, 10.0, 30.0]$.

distributional characteristics are distorted. Moreover, these synthetic datasets achieve this level of utility while also protecting the privacy of individuals, at least from attribute inference attacks to the two sensitive variables examined here (where data-driven intruder attacks did not perform substantially better than guessing based on prior information only). They also seem to offer some protection against membership inference attacks (where intruders' precision scores where close to random guess), although intruders seem to be able to pick up some information from looking at the released data; this information is unlikely to be useful though, unless intruders have access to large sets of real records beforehand. It seems that synthetic algorithms can combine utility and privacy better than the anonymisation method used in these experiments, although it remains to be seen whether this also applies to the anonymisation algorithms used by CCHIC currently. The anonymisation algorithm is particularly disclosive when faced with attribute inference attacks.

Regarding the interpretation of these results by data owners and the public, the privacy metrics used here for intruder attacks (accuracy, precision, recall, F1) are easier to explain to non-expert than concepts like differential privacy. These or similar metrics could be used in the future to allow for better transparency into the privacy protection afforded by synthetic data, especially if they are selected as a mechanism for data release of healthcare data like CCHIC's, where patient and public inclusion and consent are crucial.

The experiments in this report do not capture various types of utility that might be relevant for researchers and make assumptions about intruder knowledge that could be revised and expanded further; they also use only a small subset of the CCHIC variables. While this is an initial investigation, the results are encouraging and show that synthetic data are at least a promising direction of development for the CCHIC data release process.

9 Future work

The work presented in this report is an initial investigation on the feasibility of using synthetic data as a mechanism for releasing CCHIC data more widely. There are a number of directions that future work could follow.

This report tests two types of intruder attacks that make specific assumptions about the knowledge of the intruder and the way they perform the attack. More intruder attack scenarios could be designed and tested in the future. For example, the intruder can be assumed to hold other public information like public hospital data, access to the electoral register and the CCHIC public anonymised dataset. This would require further understanding of the data available in the public domain and what the intruder motivations might be. The latter point could be explored through the design of specific intruder types (e.g. journalist, neighbour) with different motivations, e.g. demonstrate the weak privacy protection of a dataset or disclosing information about a vulnerable individual. Additionally, some of the assumptions made in the current report might be over-optimistic for intruders' knowledge, e.g. the possession of sample datasets and the assumption that intruders know a complete record when performing membership attacks.

It would be very interesting to run the same pipeline for other synthetic methods, including deep learning generative models. These have different characteristics to PrivBayes and their differences in the quality of output are not fully understood yet. The hyperparameters of PrivBayes should also be further explored as in this report we only ran it with a specific set of them. More utility metrics should be added to this analysis to capture more use scenarios. For example, measuring correlations and correlation-like quantities for categorical variables would be useful. Finally, the actual anonymisation algorithm of CCHIC could be added to the analysis and more hyperparameter combinations could be tested for the algorithm used here.

Looking at the results of the intruder attack experiments, it is clear that the success rate of intruders for some attacks is low but also has considerable variance. It would be valuable to further understand how the intruder would choose which of their predictions they would prioritise and consider more reliable. And if there is a way to do that, would it lead to higher success rates for their attacks? There is some connection between this and the discussion about intruder motivations and what they are after. One way in which the intruder might try to increase their effectiveness is by choosing a different threshold when running their classifiers (either the classifier that predicts a sensitive attribute or the distinguisher in membership attacks). By placing the threshold higher than 0.5 (which is the value we used), the intruder might be able to make more confident guesses and increase their precision score. An analysis of ROC-AUC curves is an obvious direction for future research.

Finally, there is scope for exploring alternative privacy metrics for CCHIC and other data releases. One of them could be a version of k-anonymity that could be usefully applied to synthetic data. This could be interpretable and facilitate comparison with existing anonymisation approaches.

One of the most valuable elements of CCHIC is the rich time series data it contains which capture hundreds of measurements that happen in an ICU throughout a patient's stay. These have not been examined here but are a promising direction of research, particularly given the increasing interest in recent literature about synthetic methods for time-series data [Yoon et al., 2019b].

References

Khaled Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PloS one*, 6:e28071, 12 2011. doi:10.1371/journal.pone.0028071.

Jane Henriksen-Bulmer and Sheridan Jeary. Re-identification attacks—a systematic literature review. *International Journal of Information Management*, 36(6, Part B):1184–1192, 2016. ISSN 0268-4012. doi:https://doi.org/10.1016/j.ijinfomgt.2016.08.002. URL `https://www.sciencedirect.com/science/article/pii/S0268401215301262`.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data – anonymisation groundhog day, 2021.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019a. URL `https://openreview.net/forum?id=S1zk9iRqF7`.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4), October 2017a. ISSN 0362-5915. doi:10.1145/3134428. URL https://doi.org/10.1145/3134428.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. PrivBayes: private data release via bayesian networks. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1423–1434, New York, NY, USA, June 2014. Association for Computing Machinery. ISBN 978-1-4503-2376-5. doi:10.1145/2588555.2588573. URL https://doi.org/10.1145/2588555.2588573.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. PrivBayes: Private Data Release via Bayesian Networks. *ACM Trans. Database Syst.*, 42 (4):25:1–25:41, October 2017b. ISSN 0362-5915. doi:10.1145/3134428.

Haoyue Ping, Julia Stoyanovich, and Bill Howe. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM '17, pages 1–5, New York, NY, USA, June 2017. Association for Computing Machinery. ISBN 978-1-4503-5282-6. doi:10.1145/3085504.3091117. URL https://doi.org/10.1145/3085504.3091117.

Isabel Wagner and David Eckhoff. Technical privacy metrics: A systematic survey. *ACM Comput. Surv.*, 51(3), June 2018. ISSN 0360-0300. doi:10.1145/3168389. URL https://doi.org/10.1145/3168389.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Keerthana Rajendran, Manoj Jayabalan, and Muhammad Ehsan Rana. A study on k-anonymity, l-diversity, and t-closeness techniques focusing medical data. 17, 12 2017.

Jennifer Taub, Mark Elliot, and Joseph Sakshaug. The impact of synthetic data generation on data utility with application to the 1991 uk samples of anonymised records. *Transactions on Data Privacy*, 13:1–23, 01 2020.

NHS England's A&E Synthetic Data. https://data.england.nhs.uk/dataset/a-e-synthetic-data.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper/2019/file/c9efe5f26cd17ba6216bbe2a7d26d490-Paper.pdf.