PrivacyHub

by DATAVANT

HIPAA COMPLIANCE:
SYNTHETIC DATA PRIVACY ASSESSMENT REPORT

Prepared for

Subsalt

For its **'Synthetic Data Generation Tool'**

Prepared by

Dermot Mc Ateer PhD, MASt,

Anca Ionescu MSc, BSc (Hons),

Kristin Lund PhD, MPhys,

& Kyle McLean PhD, MSci, DIC

all of Privacy Hub

May 2024 (Not Yet Finalized)

Report: SUB22DP1a

Version History

The table below lists previous versions of this report, and states how each differs from the previous. The existence of a more recent report than this (identified by the same number but with a subsequent suffixed letter) renders this report obsolete and invalid.

Date	Report no.	Changes from previous
May 2024	SUB22DP1b	Language describing recommendations for data use agreements involved in the process. Guidance on outlier equivalence classes and language relating to continuous variables amended to align with report SUB241P1a.
December 2022	SUB22DP1a	Original report.

Contents

Executive Summary

Synthetic data is information that is based on the properties of the 'real' data and which maintains a high degree of statistical fidelity to the underlying source. In this manner, the approach of using synthetic data seeks to minimize disclosure risk while simultaneously retaining high utility in the data. Indeed, well-constructed synthetic data will commonly retain higher utility than techniques which rely on redaction or similar approaches, while also reducing the disclosure risk relative to those techniques. The process assessed in this document, and owned by Subsalt, is such a synthetic data generation process. This report seeks to assess the extent to which disclosure risk is minimized by the use of Subsalt's synthetic data generation tool as well as commenting on the enhanced privacy protection it offers in relation to and beyond the HIPAA Privacy Rule. Furthermore, the report provides evidence concerning the effectiveness of key metrics, described and tested herein, in assessing the disclosure risk of synthetic data.

Privacy Hub has closely considered Subsalt's synthetic data engine process for the generation of synthetic data from original (real) patient data, under the assumption that this original data will contain protected health information (PHI). Subsalt's current process of removing patient direct identifiers (such as names, addresses, social security number, etc.) and unique identifiers or tokens, both encrypted and otherwise, together with additional pre- and post-processing requirements introduced by Privacy Hub (which Subsalt has confirmed will be implemented), is considered effective in increasing the privacy-preserving nature of Subsalt's synthetic data generation process. Furthermore, a series of tests, concerning different types of attack on synthetic data (including attribute inference, membership inference and distance to closest record), were applied by Privacy Hub to the engine, with each attack type being tested on multiple datasets. These tests provided substantial evidence that the metrics effectively capture the level of inherent disclosure risk. In summary, Privacy Hub has investigated both the process and mechanisms used by Subsalt, as well as the synthetic data generated by the engine. Our analysis has included both examination of principles used to perform the synthetic data generation, and testing of the synthetic output to assess the associated disclosure risk to the underlying real data.

Privacy Hub is satisfied that, provided the conditions summarized in Section 4.2 continue

to be met, the process employed by Subsalt provides rigorous and robust privacy protection to individuals within the original dataset, and concludes that the synthetic data generated by this process is compatible with the HIPAA Privacy Rule. It is Privacy Hub's opinion that the tests and associated metrics for disclosure risk defined herein provide a robust mechanism to assess disclosure risk for synthetic data. However, each of the outputs of this synthetic engine (the synthetic data) should be reviewed, as detailed herein, in order to ensure that the risk of re-identification of (sensitive) information disclosure from the underlying real records remains sufficiently small.

This report remains valid until December 31, 2024, provided the conditions stated herein continue to be met. In particular, should the configuration of the synthetic engine and the models employed (as defined herein) materially, a further review of these updated models will be required to ensure continued compliance. Furthermore, the scope of this certification is limited to the assessment of the three models detailed in the subsequent sections; a review of any new models will be required to ensure continued compliance. It is advised that Privacy Hub and Subsalt convene periodically (e.g., annually) to evaluate whether technological, industry and/or academic developments have necessitated changes to the information relayed herein or to the metrics and conditions of this report. Additionally, Privacy Hub strongly recommends that a review analysis is conducted every six months, involving the assessment of a production-level synthetic dataset.

We strongly advise that this report is read in its entirety. While this Executive Summary contains the most essential points, it inherently omits detail that is still important, found within the body of the report.

Jamie Blackport	Colin Moffatt	Dermot Mc Ateer
President	Chief Data Scientist & Privacy Expert	Senior Data Scientist & Privacy Expert

all of Privacy Hub, January 2023.

Disclaimer

This report is conditional upon the accuracy, completeness and representativeness of the information supplied to Privacy Hub by Subsalt. This synthetic data privacy assessment focuses solely on the process and datasets described herein. However, we have not verified all the provided information relied upon within this report, and instead have relied upon and continue to rely upon Subsalt to inform Privacy Hub of the inclusion of any erroneous and inaccurate information, in order that changes can be made and the report remains valid. Further conditions are set out in Section 2.4, which must all be met for this report to be considered valid.

1 Introduction

This report documents the process and technology owned by Subsalt for generating synthetic data, based on information supplied by Subsalt to Privacy Hub. It assesses whether there is a risk of re-identification of individuals in the source data, or of gaining (additional) sensitive information from the synthetic data following the generation process.

The disclosure risk is assessed, building on Privacy Hub's knowledge and experience of Expert Determination under the Health Insurance Portability and Accountability Act of 1996 (HIPAA). However, the scope of this assessment is broader, and relates not only to the risk of patient re-identification, but to the risk of information gain, assessing to what extent the synthetically generated data reveals sensitive information about the individuals whose information is part of the original dataset.

Within this report, we first provide a background on disclosure risk and its relevance to both HIPAA and synthetic data. Also provided are definitions used throughout the report, and information regarding the scope of the assessment and its applicability. The suitability of the process utilized by Subsalt is then discussed, with an overview of the process provided. This is followed by analysis of the disclosure risk posed through use of the process and the generation of synthetic data. The report concludes with a concise summary of the findings and a statement on the suitability of Subsalt's synthetic data generation process with regard to HIPAA and disclosure risk.

2 Background and Context

2.1 HIPAA Requirements

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule permits the use of health information if it is not individually identifiable. It can be designated as not individually identifiable by one of two methods. This is set out in 45 CFR § 164.514, where the two methods are informally referred to as *(b)(1)* Expert Determination and *(b)(2)* Safe Harbor. The text relevent to Expert Determination is as follows:

> ***(b) Implementation specifications: Requirements for de-identification of protected health information.***
>
> *A covered entity may determine that health information is not individually identifiable health information only if:*
>
> ***(1)*** *A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:*
>
> ***(i)*** *Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and*
>
> ***(ii)*** *Documents the methods and results of the analysis that justify such determination*

While the assessment conducted herein is not strictly Expert Determination by this definition, the underlying premise is used as context for the assessment. Privacy Hub uses its knowledge and understanding of the Expert Determination process to assess the disclosure risk associated with Subsalt's synthetic engine. Further details can be found in Section 2.2.4.

2.2 Assessing Disclosure Risk

2.2.1 Disclosure Risk and Value Rarity

This section and the one following give some context around disclosure risk control and some general considerations.

Disclosure risk can be broadly categorized into two areas: identity disclosure, and information disclosure. Identity disclosure risk is concerned with how likely it is that an individual within the dataset can be identified. 'Identification' may not necessarily require matching information to a person's name, but rather makes an individual person discernible. Information disclosure refers to the risk that a patient's sensitive information is learned by an attacker aiming to derive insights from the data, and is not necessarily only about an individual. The two classes of disclosure risk are intricately linked – identity disclosure often leads to information disclosure, and information disclosure can increase the likelihood of identity disclosure. Any piece of information about an individual that is in a dataset and also in the public domain may be useful in identifying that individual, with the usefulness of that information proportionate to the rarity of the value in the population. The fewer people that share a value, the rarer it is, and the higher the disclosure risk contributed by that value. In fact, many of the steps taken in reducing risk effectively remove rare values by one means or another. The disclosure risk for an individual is proportional to the rarity of values in combination, and it is natural that individuals show variation of disclosure risk.

2.2.2 Risk and Utility

Processes which result in de-identification and risk reduction by modifying rare values in a dataset inherently detract from the utility of that dataset through loss of information. It is important to recognize that a trade-off exists between a dataset's utility and its level of anonymity; as disclosure risk is reduced, so is utility. In relation to statistical disclosure, the (Federal Committee on Statistical Methodology, 2005) states that:

"...private information organizations involved with health care data must bal-

ance two objectives: to provide useful statistical information to data users, and to assure that . . . individuals are protected."

The removal of risk from a dataset will likely remove utility, so a balance must be struck whereby the dataset retains as much utility as is reasonable while satisfying the requirements of the HIPAA Privacy Rule.

2.2.3 Synthetic Data

Synthetic data is information that has been generated based on the properties of the original 'real' data, and which maintains a high degree of statistical fidelity to this underlying data. In this manner, the approach of using synthetic data seeks to minimize disclosure risk while simultaneously retaining high utility in the data, allowing the user to largely depart from the usual risk–utility trade-off discussed in the previous section. Indeed, well-constructed synthetic data will commonly retain higher utility than techniques relying on redaction or similar approaches, while also reducing the disclosure risk relative to those techniques.

An effective process for synthetic data generation, therefore, acts to minimize the disclosure risk while retaining the data's utility for legitimate use cases. The process being assessed in this document, and owned by Subsalt, is such a synthetic data generation process. This report assesses the extent to which disclosure risk is minimized by the use of Subsalt's synthetic data generation engine, and comments on whether this aligns with the HIPAA Privacy Rule.

2.2.4 Process Assessment

The process assessed and documented within this report is evaluated in accordance with the HIPAA regulatory framework. While HIPAA (§ 164.514 (b)) makes reference only to assessment of health *information*, this report is centered around assessment of a *process*. Therefore, as an Expert Determination under HIPAA cannot verify that a process itself is aligned to HIPAA, the approach used for this report was to assess whether the process

is suitable in ensuring that all synthetic datasets generated would be considered as de-identified in relation to HIPAA, subject to the conditions stipulated herein.

This assessment of the disclosure risk associated with Subsalt's synthetic engine is informed by Privacy Hub's knowledge and experience in Expert Determination. The author(s) of this report is (are) the person(s) with appropriate knowledge and experience, the analyses and assessment documented are as described in *(b)(1)(i)*, and this report is as referred to in *(b)(1)(ii)*. However, the scope of this assessment is extended to include analyses of the risk relating to information disclosure, such that the overall disclosure risk (as defined in Section 2.2.1) is determined to be 'very small' against the HIPAA standard (as described in Section 2.1). Therefore, this report includes analyses in line with Expert Determination (as described herein) and also contains:

- Assessment of key areas within the process where the disclosure risk is raised.
- Requirements and recommendations for modifications or actions to reduce the risk of re-identification or information disclosure (where any are necessary).

Additionally, HIPAA does not quantify the level of acceptable risk, but specifies it should be 'very small'. 'Very small' is open to interpretation. Furthermore, as the emergence of high-quality synthetic data is relatively novel and there is a scarcity of literature against which privacy standards for such data may be set, rigorous thresholds are more difficult to define in relation to synthetic data.

2.3 Definitions

Here, we set out definitions of terms used within the report, where either their precise meaning is important, or their interpretation can vary.

2.3.1 HIPAA-Defined Entities and Information

We include the following definitions from Subpart A: General Provisions 45 CFR § 160.103, which are adhered to by Privacy Hub when these terms are used within the report.

Health Care Provider

Health care provider means a provider of services (as defined in section 1861(u) of the Act, 42 U.S.C. 1395x(u)), a provider of medical or health services (as defined in section 1861(s) of the Act, 42 U.S.C. 1395x(s)), and any other person or organization who furnishes, bills, or is paid for health care in the normal course of business.

Health Information

Health information means any information, including genetic information, whether oral or recorded in any form or medium, that: (1) is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and (2) relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual.

Individually Identifiable Health Information

Individually identifiable health information is information that is a subset of health information, including demographic information collected from an individual, and: (1) is created or received by a health care provider, health plan, employer, or health care clearinghouse; and (2) relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual; and (i) that identifies the individual; or (ii) with respect to which there is a reasonable basis to believe the information can be used to identify the individual.

Protected Health Information (PHI)

Protected health information is individually identifiable health information: (1) except as provided in paragraph (2) of this definition, that is: (i) transmitted by electronic media; (ii) maintained in electronic media; or (iii) transmitted or maintained in any other form or medium. (2) PHI excludes individually identifiable health information: (i) in education records covered by the Family Educational Rights and Privacy Act, as amended, 20 U.S.C. 1232g; (ii) in records described at 20 U.S.C. 1232g(a)(4)(B)(iv); (iii) in employment records held by a covered entity in its role as employer; and (iv) regarding a person who has been deceased for more

than 50 years.

Anticipated Recipient
HIPAA (45 CFR § 164.514) defines the accepted disclosure risk in relation to an *anticipated recipient*, but makes no formal definition of this term. Privacy Hub considers the anticipated recipient to be any person or persons to whom Subsalt intentionally forwards the data. (Contrast this with the definition of *Naïve Intruder* in Section 2.3.3).

2.3.2 Data-related Definitions

Synthetic data is information that is artificially created rather than generated by real-world events or direct measurement. Synthetic data is algorithmically generated; either independently or as a derivation of an underlying real dataset.

Health data comprises a number of variables (also known as features or attributes) typically arranged as columns, which can be classified according to their content and potential for use in re-identification. An understanding of how these variables differ is useful in understanding the analysis. Privacy Hub adopts the following classification:

Direct identifiers unambiguously identify individuals or small groups of individuals. Such variables include names, addresses, social security numbers, telephone numbers.

Indirect identifiers are those variables that in combination can be linked to external information to potentially re-identify individuals. They include those variables which may be readily observable, such as gender or race. On their own they may pose no risk, but in concert they may identify either individuals or small groups of individuals. As such, these variables are potentially of high risk. Such variables may also be called *key variables* or *quasi-identifiers*.

The distinction between direct and indirect identifiers is, in reality, not clear. Risk is better considered as being related to the values within a variable rather than the variables themselves. For example, the most common name in the USA will be shared by tens of

thousands of people but is regarded as a direct identifier. Despite this, we find these terms useful by virtue of their simplicity.

> **Sensitive variables** are those which constitute PHI when linked to direct identifiers, or indirect identifiers in unique or 'small-group' combinations. The disclosure of PHI may have legal and/or ethical implications. This information is often the primary reason for the existence of the dataset, and may include, e.g., the status of the individual in terms of disease incidence, health condition or therapy regime. It may also include other, less sensitive clinical attributes such as blood pressure or vital capacity. Such information is usually not directly observable and is known to a much smaller pool of people – often health care professionals – and so these variables constitute a lower risk in terms of use in re-identification; however, they can be targeted by information disclosure attacks.

A further distinction in variable type defines how they are recorded, which in turn dictates the methods appropriate for their analysis:

> **Categorical variables** are those with a limited number of possible values where intermediates are not possible. As such, these variables reflect membership of a distinct group; for example, gender, race and medical center fit this definition.
>
> **Continuous variables** are represented by numbers on varying scales. They include measurements such as height, weight and blood pressure, but also counts such as number of offspring.

2.3.3 Additional Definitions

> **Naïve Intruder**
> We use the term naïve intruder to mean an unauthorized person who has gained access to the data but has no additional knowledge to aid interpretation of those variables and values which are not self-evident. (Contrast this with the definition of *Anticipated Recipient* in Section 2.3.1.) However, throughout the assessment of Subsalt's synthetic engine, we assume that the attacker can be not only a naïve intruder but also an anticipated recipient or any person or entity that has by any

means gained access to the synthetic data produced by the engine.

Contextual Health Information

We define contextual health information as information where individuals share a common value for a health-related variable, where that variable does not occur explicitly in the dataset. Thus, such commonality may reveal sensitive health information about all the individuals within the dataset without explicitly being contained within a variable in the dataset. This contextual information can be, but is not exclusively, attached to the dataset in the form of a header or title of the files delivered. For example, a single variable of patient ID numbers may be considered contextual health information if the data file were named *Diabetes_Patients.csv*.

2.4 Scope of the Assessment

The quality of the assessment carried out by Privacy Hub is conditional on a number of assumptions which are set out below. These have consequences for the validity of the findings of this report, and the way in which the process and subsequent data can be permissibly used.

2.4.1 Accuracy of Provided Information

All information supplied by Subsalt to Privacy Hub, related directly or indirectly to the process assessed, and any subsequent datasets or tools examined, are assumed to be accurate and correct. Where diagrams or descriptions of subparts of the process are included, we rely on the client informing us of any inaccuracy or error, to ensure that the report descriptions and diagrams are accurate before the report is finalized. The findings of this report are conditional upon the accuracy of this information (as it is presented within), and should it be found that this information is not accurate and correct, Privacy Hub reserves the right to pronounce this report invalid.

It is also assumed that the process and any specific configuration implemented in the assessed version of the engine (as provided to Privacy Hub by Subsalt) are consistent with that implemented by Subsalt in its available product(s). Should the process or

configuration meaningfully alter from that assessed herein, then the conclusions drawn in this report are unsafe, and the report will, therefore, become invalid. Privacy Hub strongly recommends that any new or meaningfully updated process is independently assessed.

2.4.2 Data Privacy

This report and the process it documents are primarily concerned with the HIPAA Privacy Rule. The risk of disclosure considered within this report is in relation to data and datasets present within the process; the variables within those datasets and values therein. Where data storage environments and different entities are present in the process, if there is no mention in this report of technical and organizational methods being employed, then for the purpose of the report we make the assumption that the organization employs adequate security measures to protect any data used in the process. This report is not a consideration of the security environment surrounding the dataset; its access, storage, transfer, management and administration are covered by the HIPAA Security Rule. Ongoing security appraisals are recommended by HIPAA. These can be completed internally or by external organizations that offer this service. This report also restricts itself to assessing the privacy-related aspects of the Subsalt process, making only brief, informative comment on the utility and statistical fidelity of the engine, where relevant; it does not seek to provide a thorough assessment of the engine in this regard.

2.5 Approach to Analysis

2.5.1 Disclosure Risk from Synthetic Data

Prior to discussing the approaches used in this analysis, it is worth briefly restating the forms in which disclosure risk could be present in *synthetic* data, as these can differ substantially from the ways in disclosure risk is manifested in *real* data.

While real data poses a risk of directly revealing individuals' values to a potential attacker, the risk posed by attack against synthetic data is that the attacker could use the synthetic

information to determine whether specific individuals are part of the underlying original dataset, or to deduce certain (sensitive) characteristics of the real records. The key tasks of a synthetic engine in a privacy context could, therefore, be considered as follows:

- To construct the synthetic data in such a manner that an attacker cannot determine whether an individual with those characteristics exists in the original dataset, and
- To ensure no additional sensitive characteristics of an individual within the real dataset can be inferred by an attack against the synthetic dataset.

2.5.2 Implications for Approach

As disclosure risk is evidently impacted by the types of information captured within a dataset, it is of paramount importance to consider how datasets formed from different types of variables perform under specific disclosure attacks. As such, datasets containing a variety of variable types were used in this assessment in order to provide a thorough review.

Synthetic data is designed to be flexible to the customer and to handle a wide range of data. For instance, synthetic data may be generated from variables which could range from those posing minimal risk of disclosure (e.g., the result of a coin flip carried out by a patient), to those posing a high degree of risk (e.g., the demographic information of a centenarian patient). As the engine should, therefore, be designed to be resilient against attack for all forms of data, this assessment focuses on the principle that different forms of data could be attacked. However, there are certain variable types whose exclusion in the pre-processing phase is required by Privacy Hub. This is discussed in more detail in Section 3.2.

Secondly, to maximize Privacy Hub's ability to gain personal information from the synthetic engine, Subsalt has also provided precise details of the technology and algorithms employed in the construction of synthetic data. This approach is in accordance with Kerchoffs' principle and Shannon's maxim (Shannon, 1949) (commonly used in cryptography), which states that *"one ought to design systems under the assumption that the enemy will immediately gain full familiarity with them"*. Even though it is not anticipated that a

given attacker would have access to this information (and, therefore, be able to *"gain full familiarity"*), Privacy Hub's access to it allows the construction of attacks which have the maximum likelihood of success and, therefore, provide the most conservative assessment of the disclosure risk.

Additionally, in assessing the inherent disclosure risk, Privacy Hub does not rely on the condition of the naïve intruder as defined in Section 2.3.3; rather, it assumes the presence of an educated attacker with knowledge of the principles of synthetic generation – an attacker who is able to construct sophisticated methods of attack. This is despite the fact that all engine configuration is completed by Subsalt, and Subsalt's clients are only provided with the synthetic data.

Subsalt will have data use agreements (DUAs) in place with its clients to further mitigate any disclosure risk associated with the sharing of synthetic datasets. Throughout Section 3, this report highlights a number of features that are required or recommended to be included in these agreements. Further to this, Subsalt's clients (the "data owners") are expected to share this data with their own clients (the "data consumers"). DUAs must equally exist between the data owners and the data consumers such that all protections highlighted in this report are maintained. Subsalt will advise the data owners of the relevant conditions required to protect the privacy of the synthetic data to ensure protections are propagated to future recipients of the data.

The above principles being considered, it is recommended that a statistical disclosure risk certification is updated periodically as factors in the external environment such as technology, social conditions, and the availability of information change over time. Indeed, as the generation of synthetic data is a newly developing field, further advances in industry understanding and technology mean that the advice given in this report may benefit from regular review, wherein, if applicable, updates to any and all requirements, recommendations and quantitative thresholds can be made.

3 Analysis

In this section, we present a description and assessment of Subsalt's synthetic data generation process at a level deemed sufficient to consider the associated disclosure risk in the context of HIPAA regulations. It should be noted, however, that some elements of the process are proprietary. While these elements have been made known to Privacy Hub and have been included in the assessment, they are not detailed within the body of the report, in order to allow Subsalt to maintain protection of intellectual property.

3.1 Description of Synthetic Generation Process

Subsalt's synthetic generation process can be divided into five distinct steps:

1. Ingestion of data
2. Pre-processing
3. Synthetic generation
4. Post-processing
5. Query access

We discuss each step in more detail below, making reference to each with the number shown above.

1. Ingestion of data
In step 1, the client data is ingested into Subsalt's process. Delivery of the data falls under the HIPAA Security Rule and Privacy Hub makes no statement on the security of the transfer of files between Subsalt and its clients, instead assuming that Subsalt has adequate safeguards in place in this regard. Subsalt has the option of either using its cloud-hosted environment or working within a client's virtual private cloud (VPC), depending on the security needs of the client, and where the data is physically delivered is dependent on these requirements.

During this stage of the process, decisions will be made by the client with regard to how the synthetic output should be formatted. Where relevant, Subsalt will be involved in this part of the process. This decision will be informed by the use case and structure of the real data, including the expected distributions or characteristics of the individuals within the dataset. Decisions are made on which variables to include and exclude and on the size of the synthetic outputs. While it is general practice that a synthetic dataset of similar size to the real dataset is produced, optionally, a synthetic dataset of a different size can be created. Subsalt has the capacity to train a number of different configurations of each model type, generating synthetic outputs for each. Therefore, it is likely that a number of datasets are generated from the same real data.

Repeated synthetic generation based upon the same (or similar) real dataset(s) can lead to a deeper understanding of the inner workings of a synthetic engine, leaving it more vulnerable to attack. With access to a large number of synthetic datasets, each generated from the same real dataset, an attacker may learn, for example, the distribution of values contained within a particular numeric variable, allowing inference of local and global maximum or minimum values (see Section 3.2.6 for a discussion of such values). Additionally, generating a number of synthetic datasets from the same real dataset may increase the chance i) that a real record is reproduced within the synthetic output (as discussed in Section 3.7), and ii) of accurately inferring that a particular individual was part of the real dataset (as discussed in Section 3.9). With an additional understanding, if gained, of the particular models used to generate the synthetic outputs, the particular strengths and weaknesses of the models may be used as a possible route for an inference attack.

Subsalt currently places a limit on the number of datasets that can be generated from a single underlying dataset. Currently, synthetic datasets are only generated using configurations for three models: TVAE, CTGAN and CopulaGAN. However, any limitation on the number of datasets generated from each model (with varying configurations) is currently based on resourcing (as the generation of large numbers of separate synthetic datasets places an impractical burden on Subsalt's resources); there is no quantitatively defined limit. Subsalt will have data use agreements (DUAs) in place that limit the number of synthetic datasets available to users at any given time, thereby ensuring that

statistical inferences cannot be made against the synthetic engine and the associated synthetic output. Additionally, Subsalt has confirmed that it will have DUAs in place with its clients, ensuring details of the models used for generation are not shared downstream and emphasizing the need to ensure continued compliance in this regard. Indeed, Subsalt has informed Privacy Hub that, generally, it does not not divulge these details.

It is expected that in order to make accurate inferences under such attacks, hundreds or indeed thousands of datasets would need to be generated from the underlying real data. Therefore, a sensible level of monitoring in this regard should suffice to ensure that the risk of inference remains low. Provided Subsalt continues to do this, no further action is required, though requirements related to data that forms part of a recurring feed are detailed in Section 3.6. Additionally, Privacy Hub recommends, and indeed assumes, that in order to mitigate this risk when sharing synthetic datasets created from the same real dataset with *different* clients, a DUA will be in place between Subsalt and any individual recipient of the synthetic output. This will ensure that synthetic data generated on the basis of the same real dataset cannot be shared between multiple recipients, thereby significantly reducing the risk that an attacker may successfully gain unregulated access to several synthetic datasets from different sources.

On completion of step 1, the process of data preparation begins.

2. Pre-processing

In step 2, Subsalt engages in the pre-processing stage, during which the real data is processed before it is used for synthetic generation. Whether or not the incoming real data has undergone some level of de-identification, Privacy Hub does not assume, *a priori*, that any specific pre-processing is carried out by Subsalt, and considers the requirements and recommendations detailed herein sufficient to ensure that patient privacy is suitably protected. The pre-processing stage includes, for example, the removal of information which is not required for synthetic generation. In particular, certain variable types which are currently deemed to carry too much risk are removed, such as direct identifiers, in addition to those which will add no utility value to a synthetically generated dataset, such as unique identifiers. Additionally, those variable types which are not fully supported in the synthetic generation process are removed during this pre-processing stage. This

currently includes free-text variables, which fall outside Subsalt's standard process of synthetic generation. More details on free-text variables can be found in Section 3.2.15.

3. Synthetic generation

In step 3, the synthetic data is generated using Subsalt's proprietary algorithm, building on open-source libraries. Details of the process by which the synthetic data is generated are not made available to Subsalt's clients. Privacy Hub has not conducted a thorough review of the engine's system architecture; therefore, while Subsalt has provided descriptions of the engine and, where relevant, specific references to the architecture are made throughout this report, to some degree the algorithms and models used are considered a 'black box' for the purpose of this review. In particular, this review examines how the engine performs in terms of the production of synthetic data, and focuses primarily on the privacy aspects of the synthetic output. It does not make comment upon the suitability of the model's architecture in generating this data. We will refer collectively to the process and algorithms used to generate the synthetic data as the 'synthetic engine' (or similar) throughout the remainder of this document.

Once data is received, Subsalt trains a number of models, producing synthetic versions for each model. Each model will produce a separate synthetic output based upon the same underlying real data; therefore, the number of models should be kept sufficiently small such that statistical inferences can not be made by comparing these outputs. See Section 3.5 for further details. Certain models are more suitable than others for particular types of input data, depending on the structure and distributions involved, and different models involve different levels of configuration. Currently, Subsalt trains three models for production-level synthetic data. These three models only, and their configuration by Subsalt, are the subject of this review. Other models used for synthetic generation lie outside the scope of this certification. See Section 3.2.11 for more information.

4. Post-processing

Following the production of a synthetic version of the data, there is a post-processing stage where a number of privacy and utility metrics are computed. If any of the privacy metrics discussed within subsequent sections of this report fail, as defined by the thresholds set out herein, then further discussion occurs between Subsalt and its client and a decision is

reached on how best to proceed. Further information on the disclosure metrics and the associated disclosure risk thresholds can be found throughout this report. Similarly, utility metrics measure how well the synthetic data captures the statistical relationships present within the real data. While this report makes a few minor comments on the statistical fidelity of Subsalt's synthetic engine, it does not constitute a thorough assessment in this regard; rather, it focuses on measuring the privacy of the synthetic output in order to ensure that it is deemed suitably de-identified against the HIPAA standard.

5. Query access

Once the prior steps have been completed, the synthetic data is ready for delivery. At this stage, Subsalt saves and stores the final model configuration for the approved synthetic output. The data itself is removed from the environment, along with the real data. Subsalt does not store either of the datasets. Rather, the model paramaters are stored such that the same synthetic output can be generated within the client's preferred environment on request, using deterministic sampling. The generation can occur in a client's VPC or in Subsalt's cloud environment, depending on the security requirements. However, it is important to note that all details relating to the definition and configuration of the models themselves are controlled by Subsalt, and Subsalt has confirmed that clients will at no stage have access to this information. Rather, clients will have the ability to submit SQL queries against the trained models via a user interface, from which they can return the synthetic datasets or subsets thereof. Clients may also detail some additional column-class definitions and constraints (see Section 3.2.9). Once a query is submitted to the environment, a number of models may be trained on the data producing synthetic output. At no stage will a user know which model has been used for training.

It is important to note that these queries are not generating new synthetic data. The synthetic outputs are defined in the previous steps. All queries will return results from one of the predefined model configurations. Similarly, whilst a client may submit the same query multiple times, each query is directed to a previously defined synthetic output – i.e., the query itself does not generate any new synthetic data. This is an important feature from a privacy perspective as it is not possible to generate a large number of synthetic datasets from a single repeated query, thereby limiting the ability to gain information via statistical inference over large generations. There is a finite cap on the number of model

configurations available for querying, and although there is currently no defined hard cap on the number of such models, it is currently limited to configurations based upon three models and it is expected that the number of models will not rise significantly. Any new models, not detailed herein lie outside the scope of this report. Further information can be found in Section 3.5.

3.2 Pre- and Post-Processing

In assessing the privacy of Subsalt's synthetic engine, Privacy Hub does not assume that the real input data has been de-identified; we instead assume that any variables contained within the real data may be identifying, and detail explicitly where modification to these variables is required. Indeed, Subsalt does not require that any level of de-identification is performed on the incoming data, *a priori*, before generating the synthetic output. In the following sections, we detail changes which must be made, where relevant, to the underlying real data *before* any synthetic version of the data is produced. These modifications help to ensure that the generated synthetic data carries a sufficiently low risk so as to be considered de-identified against the HIPAA standard. The entirety of these requirements does not ensure that the real input dataset has been de-identified; their sole purpose is to help ensure that any synthetically generated data carries a sufficiently low level of risk.

3.2.1 Direct Identifiers

Direct identifiers are those variables which unambiguously identify individuals or small groups of individuals. They include names, addresses, social security numbers and telephone numbers. While any direct identifiers that are processed by a synthetic engine are likely to be distorted, it is important to suitably protect against direct inference (for example, via engine memorization), or indirect inference via back-engineering, of any direct identifiers in the underlying real dataset from which the synthetic version was generated. The level of risk inherent in these identifiers is enough to require the redaction of all such variables before the real dataset is processed by the synthetic engine. Subsalt has confirmed that direct identifiers are removed by the client prior to processing by the synthetic

engine; provided this is the case, there is no further requirement to modify such variables.

3.2.2 Unique IDs

Variables containing unique tokens which identify individual patients are generally unintelligible to a naïve intruder, particularly if they have undergone some level of encryption, as they convey no information about the patient. However, the presence of these identifiers introduces a measure of risk by allowing the anticipated recipient to link to other data that possesses the same values. Privacy Hub assumes that no such identifiers appearing in the real data are susceptible to any form of attack or decryption that would allow elucidation or inference of any identifying information about the patient. Again, while any such tokens appearing in the synthetic output are likely distorted, the presence of these identifiers introduces a measure of risk if it is possible to infer the real value from the resulting synthetic generation. Additionally, the presence of such tokens in any synthetically generated output is likely to provide little in the way of utility. Therefore it is required that all such tokens and IDs are redacted before the real dataset is processed by the synthetic engine. Subsalt has confirmed that unique IDs are removed prior to processing by the synthetic engine; provided this is the case, there is no further requirement to modify such variables.

3.2.3 De-duplication

Duplicate records are a consideration for any project involving the analysis of data. If untreated, the presence of duplicate information can skew statistical results, leading to inaccurate conclusions. In synthetic data, concerns must be addressed both from a utility and a privacy perspective. When generating synthetic data, it is important to preserve the statistical fidelity of the underlying real dataset. However, if the real dataset is itself of poor quality, then the synthetic data, *a priori*, will likely be of little use. Therefore, in order to ensure that the synthetic generation is optimized for whatever use case it has been designed for, the underlying real data should be as clean as possible in this regard. The more information that can be garnered cleanly from the underlying real dataset, the better the synthetic generation will perform under similar analysis.

In addition to these fidelity concerns, duplicate records carry a level of privacy risk. Even on the assumption that a synthetic engine does not unduly memorize records, the goal of any synthetic engine is to produce a statistically accurate synthetic version of the underlying real dataset. Therefore, any duplication in the real dataset is likely to be seen proportionally within the synthetic dataset. This makes it easier to back-engineer records and learn how the engine behaves, both directly by learning the information contained within the duplicate records, and also indirectly via similar records, i.e., those generated from the duplicate real group but which have a degree of variation across their variables in the synthetic data. For example, if a group of duplicate real records leads to a large proportion of duplicate synthetic records, it may be possible to infer that these duplicates are in the real data. Additionally, if there is a large number of similar records, for example, differing from the duplicates in only one numeric feature, then it may be possible to infer the engine's behavior in regard to numeric features more generally, which would aid in the back-engineering of numeric features within the entire dataset. This is not usually possible, as an attacker cannot be certain of the form of the underlying real record from which a synthetic record is generated. With duplicates, however, it is possible that an attacker may be confident that such similar records are indeed generated from some known duplicate records, thereby allowing the real record to act as a control from which to analyze the similar synthetically generated records.

As previously mentioned, there is also the more direct possibility of a disclosure breach: if the duplicate records in the real data are carried through to the synthetic data, those records may be readily identifiable – particularly if, for example, the synthetic data, outside of any duplicate records, generally produces unique rows. This may allow an attacker to uniquely identify a patient along with any health information contained within the record. Further information can be found in Section 3.8. Privacy Hub requires that all duplicate records (matching across all variables) are removed from the real dataset during pre-processing, *before* generation of the synthetic data.

3.2.4 Removal of Highly Identifiable Records

Some patient records are at higher risk of being re-identified within real health data, as they exist within small groups (sharing the same indirect identifiers) within the real

population. Consider, for example, a health dataset containing the following patient indirect identifiers: age, gender and 3-digit ZIP code. The record of a 40-year-old man and a 111-year-old woman residing in the same 3-digit ZIP code will pose different levels of disclosure risk despite the same three indirect identifier types being disclosed. When looking at the real population of the 3-digit ZIP code area, it will most likely be the case that a 111-year-old individual is found in a very low proportion, while it is very likely that there is a relatively high percentage of 40-year-old individuals within the area. Thus, the ability to link patient indirect identifiers present within the given health dataset with reasonably available information escalates the disclosure risk of the second individual.

On the other hand, such high-risk records cannot simply be removed from the real dataset before synthetic data is generated, as this would unduly affect the fidelity of the synthetic data. Moreover, it may well be that the synthetic data does not even include a record that is highly similar to the original high-risk patient record, due to the statistical variation of the engine. However, there are cases in which the synthetically generated dataset may contain a record (or records) deemed so similar to the real record that it increases the associated disclosure risk. This is considered further in Sections 3.7 and 3.10.

3.2.5 Dates

Dates are first converted by the engine to a numeric Unix timestamp – a datetime representation measuring the time in seconds that has elapsed since the initial reference datetime 00:00:00 UTC on January 1, 1970. In this format, the dates are then processed by the engine as any other continuous variable. Once the synthetic value is generated, the synthetic Unix timestamp is converted again to a standard date format. There is a mechanism within the engine to ensure that future dates[1] do not appear in the synthetic output. Indeed, this mechanism ensures that all numeric fields (see Section 3.2.6), unless otherwise specified by the user, maintain the same upper and lower bounds on numeric variables as found within the real data. If values are generated which are larger (lower) than those which occur in the real data, they are set to the maximum value. This may

[1]Here we refer to those dates which lie outside the maximum date range of the input data. Where the input data includes dates in the future, relative to the latest date-time, we can expect to see such dates occurring in the synthetic output.

create a disproportionately large group of values at the maximum. Subsalt has confirmed that, as a result of the generation process, dates additional to those present within the real dataset will be present within the synthetic dataset, and that these will vary in accordance with the distribution of real values.

It is Privacy Hub's judgment that such an approach will considerably reduce the risk arising from variables containing dates, particularly those which could introduce a high level of disclosure risk if memorized from the real dataset or inferred from the synthetic dataset. Thus, Privacy Hub advocates the current approach employed by Subsalt, as detailed herein, and recommends that a review is performed if this approach changes in the future or for any special case.

From a privacy perspective, some types of dates introduce more disclosure risk than others. For example, dates of birth and dates of death are considered indirect identifiers; while, for example, a date indicating when the patient picked up a prescription is not identifying and carries minimal risk. Additionally, in instances where accurate dates of birth or death can be inferred, careful consideration must be given to the risk. For instance, a date of service in combination with a Current Procedural Terminology (CPT) code indicating that the patient is deceased (such as code '88045' — 'Necropsy (autopsy) coroner's call') will provide a close to exact date of death. Privacy Hub requires that, where relevant, inferred dates of birth and death are included in all risk calculations described within this report, including calculations in relation to the distance to closest record (DCR), attribute inference and membership inference. More details on the risk associated with matching records, and information gain related to dates of birth and/or dates of death, can be found in Sections 3.7 and 3.10.

With regard to dates of birth and death, Privacy Hub proposes that a sensible approach would be to consider a real and a synthetic record to have the same (or similar enough) such attribute (and, thus, to match on this quasi-identifier) if the following conditions are met. Firstly, for date of birth, we recommend that a real and a synthetic record are considered to match if the two values match on the calendar year. As such, in assessing the match between a real and a synthetic record, date of birth is addressed in the same way as any age variables, which Privacy Hub feels is sufficient due to way in which dates

are processed by Subsalt's engine and based on the knowledge of reasonably available information. However, more granular reasonably available information exists with regard to date of death, and can be at the level of granularity of day of death. Thus, we propose a more conservative approach when matching based on date of death variables. If the only information present within the dataset is year/month of death, then the year/month should be used to determine the match. However, if the full date of death is available, then the dates of death of a real and synthetic patient record should be considered a match if they are within the same 14-day period. These recommendations should be applied to all types of tests presented within this report when the datasets contain dates of birth/death.

3.2.6 Numeric Features

Numeric features are processed according to each model's architecture. Importantly, they are not treated as categorical; therefore, values which do not occur within the real data will be seen within the synthetic output. This will generally help to lower the disclosure risk associated with numeric variables.

Data by default is rounded to the furthest digit seen in the source data. For example, where all values in a height variable are rounded to the unit (e.g., '1', '2', '3' ...), but one record goes to the tenths (e.g., '1.2'), the engine will default to display all synthetic values rounded to one decimal place based on this single record. Users can specify the rounding granularity (see Section 3.2.9 for more information on user-defined constraints) and this does not impact synthesis. At the time of generation, the data is rounded to the granularity specified; if no rounding constraint is specified, the engine defaults to use the most granular level appearing in the data.

Furthermore, Subsalt has the ability to modify the synthetically generated values. Where deemed relevant, Subsalt can remove the top and bottom x% of values to ensure that extreme outliers in either direction are removed from the data. This will generally lower the disclosure risk in situations where it is applied. Similarly, Subsalt may choose to maintain the maximum and minimum values, as contained in the real data, throughout the generation process. Where values are generated which are greater (lower) than the maximum (minimum) value, they may be mapped to the maximum (minimum). This

means patients with outlier values in certain demographic features, relative to the real population, may be more readily identifiable – for example if a male aged 116 occurs in the real data and the age is carried through to the synthetic output. The associated risk of meaningful information gain will be calculated as part of the attribute inference analysis as detailed in Section 3.10.

3.2.7 Categorical Features

Categorical variables are processed according to each model's architecture in a manner which preserves the value range of the real data; i.e., no values will appear in the synthetic data that did not occur in the real data. Binary data is treated as categorical (it is converted to a string if in a numeric form).

Categorical fields' values are generally subject to a minimum occurrence rate. Any value occurring less frequently than this in the real data will be converted to 'subsalt_other' prior to synthesis. The minimum is set by default to 1% of the total number of rows, with a hard floor defined at 100. For example, where there are 100,000 rows in the real data, a minimum occurrence rate of 1,000 must be met. Where there are 5,000 rows in the real data, the 1% threshold would define the occurrence rate at 50 records. Therefore, the hard-floor occurrence rate of 100 would be used in this instance. However, it is possible to remove the minimum occurrence rate, where required, for certain use-cases. Whilst the presence of 'subsalt_other', if configured as detailed in Section 3.2.13, will generally decrease the privacy risk to the underlying patients, the conclusion of this report is not dependent on its use. Indeed, the majority of the analysis detailed herein was conducted on synthetic outputs in which this minimum threshold was not set. No such thresholds are defined for binary variables, where it is assumed that any such mechanism can be readily back-engineered to reveal the obscured value. For binary variables, the risk associated with these values in relation to meaningful information gain will be accounted for via attribute inference testing as detailed in Section 3.10. Privacy Hub condones the 'subsalt_other' mechanism as good practice. More details on this can be found in Section 3.2.13.

3.2.8 Computed Features

Certain variables are not processed by the synthetic engine and are instead 'computed', generally based on another variable occurring within the data. Computing variables in this way allows certain relationships to be maintained through the generation process. These include, for example, the relationship between age and year of birth. Computed fields are calculated at generation time and specified by the user. For example, where age and year of birth variables are present, year of birth can be synthesized from the real data. Once synthesis is complete, age can be calculated from the synthetic values. For non-calculated numeric data, including dates (calculated as detailed in Section 3.2.5), the maximum and minimum values are by default set to the maximum and minimum of the input data. For calculated variables, discrepancies may arise due to data quality. Users can configure custom maximum and minimum values, in which case rejection sampling is used at the time of generation to ensure the requirements are met.

Users can also specify that combinations of categorical variables in the synthetic data should only appear as they appeared in the source – for example, for ZIP and state variables. In such cases, prior to synthesis, the variables are combined into one column using a unique delimiter, '#', and it is this new variable which is processed by the synthetic engine. At time of generation, this column is split using the delimiter, creating individual variable columns matching the layout of the real data.

3.2.9 Constraints

Constraints within the Subsalt engine are deterministic rules defined by the creator of a synthetic dataset, prior to synthetic generation. For example, a user can define a constraint on variable A, such that variable A = variable B + variable C. A constraint must be true at least 99% of the time within the input data to be considered valid. If the constraint is valid, it will be enforced 100% of the time in the synthetic data. If not, it will not be applied. There are a variety of constraints that users can configure.

3.2.10 Information Regarding Patient Residency

As detailed in Section 3.2.1, direct identifiers such as addresses are removed by the client during the pre-processing stage. However, some residential facility names and identifiers, as well as certain codes indicating patient residency at a facility location (such as revenue code '0525' — 'Visit by a practitioner to a member in a residential facility'), in combination with a facility name, address or other value (or combination of values) that allows inference of the facility's address, may reveal the patient's residential address.

To minimize the risk associated with (the combination of) such codes/values being reproduced within the synthetic output, Privacy Hub requires that for those synthetic records with a DCR equal to 0 (see Section 3.7 for details on DCR), either i) all information relating to the facility location should be redacted from the synthetic patient record, or ii) the code indicating patient residency should be redacted from the synthetic patient record in the post-processing stage. Privacy Hub keeps an updated list, in spreadsheet form, of the codes within each code set which indicate residency. This can be supplied on request. In the case of an exclusively residential facility such as a prison, the name and any other facility identifiers should be redacted regardless of whether or not the code has been modified. Additionally, for synthetic records that include indirect identifiers which match those of a real record, Privacy Hub requires that the same modifications are made to the synthetic records if the matching records exist within an equivalence class containing fewer than five patients in the real data.

3.2.11 Models

There are many methods for generating synthetic data, and many types of machine learning models which can be used as the basis for synthetic generation. Subsalt's synthetic engine is built upon a library of such models. Currently, in order to produce production-level synthetic data (other models may currently be in use for testing and development), Subsalt employs three different models. In the course of this assessment, Privacy Hub has evaluated Subsalt's configuration of these three models alone, and provided the conditions detailed herein are met, this report remains valid in relation to those three models. However, other configurations (which may differ substantially from those discussed throughout

this assessment and defined herein), and indeed other models, lie outside the scope of this certification, and Subsalt should seek further assessment in this regard, where required.

3.2.12 Contextual Information

Contextual health information refers to health information that is inferred from other means than the data itself, such as a certain diagnosis being used within the name of the files sent. This specific scenario is considered further within Section 3.9. Moreover, Privacy Hub requires that where only one value populates a column within the real data, this column is discarded at the pre-processing stage. This ensures mitigation of the risk of deducing from the synthetic output that all patients share that characteristic within the real dataset. Additional details regarding the contextual information in the context of 'subsalt_other' values can be found in Section 3.2.13, below.

3.2.13 'Subsalt_other'

During pre-processing, Subsalt has the ability to define a family of parameters, which we will refer to herein as κ_i, for each variable i in the data. For each categorical variable in the real data, κ_i defines a minimum occurrence threshold of a given value within the variable. Any values with an occurrence rate falling below this value in the real data will be replaced with the value 'subsalt_other' before synthesis. This 'subsalt_other' value will then be processed by the synthetic engine as usual and will appear proportionally in the synthetic output. Whilst Subsalt can implement these thresholds on a variable-by-variable basis, generally, a single threshold is applied across the entire dataset.

The parameter has a default value of 1% with a hard floor of 100. This obfuscation technique will, in general, reduce the disclosure risk of the underlying real patients as it removes values which occur less frequently in the real data. Privacy Hub condones this as good practice; however, due care is needed when implementing such a measure. For example, consider the scenario where an attacker has access to a list of values appearing in a real dataset, perhaps through a data schema, as well as the synthetic output. Let's say an example variable 'X' contains ten categories in the real data. Let's also say that values for nine of these categories satisfy the minimum occurrence threshold. In the synthetic data,

these nine categories will appear along with the additional value 'subsalt_other' generated from records exclusively falling below the minimum occurrence threshold for variable 'X'. The attacker can readily infer the true value of variable 'X' in these instances, and can perhaps make more general inferences for the record at large. We cannot simply replace occurrences of 'subsalt_other' with a null value, as where variables otherwise contain no null values, accurate inferences may again be made. Rather, Privacy Hub requires that a new threshold is defined in conjunction with this value, such that, in scenarios where only one value falls below the κ_i in the column, the 'subsalt_other' obfuscation is not used. In these scenarios, the associated risk will be captured by attribute inference testing (see Section 3.10).

3.2.14 Missing Data

Subsalt has informed Privacy Hub that where there are missing values in the real dataset, they will be treated in the following way. Categorical variables will be treated as any other category and will appear proportionally within the synthetic output. For continuous variables, a new column will be created in the data which flags if the value in the relevant variable is null. Both columns are then processed by the synthetic engine. Once the synthetic data is generated, for all indexes in which the synthetic null flag is marked as positive, the associated synthetic value is set to null. For example, in processing a *height* variable, a new column is created in the real data – say, *height_null_flag*. The real *height* and *height_null_flag* values are synthesized, and where the synthetic *height_null_flag* variable has a positive result, the synthetic *height* value is set to null. The *height_null_flag* variable is subsequently removed. Importantly, from a privacy perspective, the missing value is not imputed by the engine.

Privacy Hub endorses this approach and considers it a reasonable method for processing missing values. As such, when there is missing data within one column of the real dataset, it is expected that missing values would appear proportionally in that column within the generated synthetic data. A specific recommendation relating to missing values can be found in Section 3.10.

3.2.15 Free-Text

Free-text variables pose a potential risk as they may contain patient identifiers, entered either deliberately or inadvertently by an operator. Free-text variables have no predefined format, and any combination of information has the potential to be available in the form of words, sentences, or numerical values which do not have limited, predetermined options or align to a known range of values. High risk lies within these variables if they are not processed to remove any identifying information. The nature of free-text variables also means it is considerably more complex to reliably determine if patient identifiers are present within them. Without a comprehensive review, quantification of the risk contained within these variables is not possible. For this reason, Privacy Hub requires either redaction of all free-text variables within the underlying real dataset prior to processing by the synthetic engine, to remove any risk contributed by the potential multitude of identifiers present; or a full assessment of the real values contained within those variables, again prior to processing by the engine, to ensure that no personally identifiable information (PII) is revealed and can be subsequently generated by the synthetic engine.

Subsalt has confirmed that it does not currently process free-text variables, and that all variables containing free-text are removed during the pre-processing stage such that no free-text is processed by the synthetic engine. Provided this is the case, there is no further requirement to modify such variables.

3.2.16 File-Naming Conventions

As per the definition of contextual health information (see Section 2.3.3), the naming convention for files delivered to and by Subsalt is a source of possible attack. For example, if Subsalt's client were to send a real dataset to Subsalt entitled 'HIV_data_1.csv', which an attacker can see, then the attacker may infer HIV status for included patients. By extension, this would apply to the synthetic output if, for example, Subsalt were to title the resulting data file 'HIV_data_1_synthetic.csv'. Therefore, due care should be taken with regard to both incoming and outgoing datasets. Subsalt should ensure that any potentially risky datasets which have been delivered by its clients are suitably renamed to remove any identifying information and, thus, to sufficiently lower the disclosure risk.

Such naming conventions also apply to those datasets delivered by Subsalt to its clients.

3.3 Minimum Dataset Size

The size of the input data has a significant bearing on the success of the synthetic engine in creating data that maintains a high level of both privacy and utility. Synthetic data, generated by the application of machine learning models to a source dataset, such as that described herein, relies on training the engine to perform on an input dataset. Both the quality and size of the dataset can affect the engine's performance. If, for example, a real dataset consisted of only ten records, it is unlikely that the machine learning algorithms would capture the data's profile with any confidence, being unable to distinguish small coincidences from real relationships. Therefore, it would probably not produce synthetic data of great quality. From a privacy perspective, the engine would be likely overfit to the training data and more likely to produce similar if not identical records in the synthetic output, thus increasing the disclosure risk of the underlying real patients. Based on testing, Privacy Hub expects that small datasets will perform very poorly against any reasonable suite of privacy-preserving testing and, as such, small datasets will likely fail at this stage. Therefore, Privacy Hub requires that very small datasets are not used for synthetic generation. Very small, here, is open to interpretation and is also dependent on the number of columns contained in the dataset. As such, Privacy Hub recommends that a real training dataset (which may represent either all or a subset of the input data) contains a minimum of 3,000 records in order to protect against the increase in disclosure risk associated with training a machine learning algorithm on a small dataset.

3.4 Linking Data

In a scenario in which two companies provide data to a third party for linking and synthetic generation, an attack may occur if one of the data owners attempts to infer information about the other dataset. This can take the form of either a membership inference attack or an attribute inference attack. More information on these attack types can be found in Sections 3.9 and 3.10, respectively. In the current scenario, during a membership inference attack, the approach taken is very similar to that described in Section 3.9, though the

attacker must differentiate, not between records that were part of the real dataset and records that were not part of the real dataset, but between records that were part of their own dataset and those that were part of the other supplied dataset. In an attribute inference attack, the attacker tries to infer information about patient records outside of their own dataset. The attack would progress in a similar manner to that described in Section 3.10. The risk associated with these attacks and any associated requirements or recommendations are included in the respective sections. Privacy Hub recognizes the potential privacy threat from such a scenario; however, we assume that where real data is provided by two different Subsalt clients for linking and generation of synthetic data, data use agreements are in place and the two (or more) clients do not try to breach the privacy of the patients within the other dataset. Moreover, Privacy Hub assumes that if this scenario were to occur in the future, Subsalt would review the additional risk that might arise from the linking of two or more datasets on an *ad hoc* basis.

3.5 General Inference Test

Another class of attack that can be conducted against a synthetic engine can be more generally referred to as inference attacks. These attacks are similar to the membership and attribute inference attacks described previously, whereby the attacker aims to infer additional information about a patient, which can in turn increase the disclosure risk or attack potential for the associated record. Such attacks can include, for example, an attempt to understand how synthetic data treats numeric values, focusing in particular on maximum and minimum values, from which it may be possible to determine the age of the oldest patient in the real data or the size of a particular subgroup contained within the bounds of local maxima and minima.

However, generally, in order to attempt this kind of attack, a large number of trials need to be conducted in order to produce statistically significant results and reach a firm conclusion. This type of attack is, by and large, more suited to synthetic engines which allow direct querying of data. Subsalt does not allow for repeat queries at this scale. The synthetic generation process is currently controlled by Subsalt and there is a limit (although not strictly defined) to the number of synthetic datasets which may be

generated at any one time and by any one client. This falls well below the level needed to conduct such tests. Additionally, as discussed in Section 3.1, Subalt will not share details of the models *a priori* with clients and will have data use agreements in place to ensure these details are not used to facilitate an attack on the synthetic data.

Subsalt does have the ability to create a single, or multiple larger dataset(s). However, Privacy Hub expects the distribution of multiple generated datasets to match that of a single larger dataset. For example, we would expect five generated datasets each consisting of 100,000 records to match the distribution of one dataset consisting of 500,000 records. Contrary to this, an attacker cannot expect to derive statistically significant results by dividing one large dataset into equally proportioned subsets and treating these subsets as individually generated synthetic datasets. Additionally, Subsalt is able to process recurring feeds of data, which can introduce the risk of making general inferences on a dataset due to a large number of repeated generations. More details on this can be found in Section 3.6.

3.6 Recurring Feed

Data forming part of a regular feed needs particular consideration both from a privacy and a utility perspective. Take, for example, the scenario where a real dataset is processed by the synthetic engine. Where the real data is then updated at a regular cadence, with each update also being submitted for synthetic generation, there are generally two options for processing:

1. At the point of each update, the entire real dataset, including the latest and all prior updates, is submitted to the synthetic engine for a full regeneration of the data.

2. Each update, alone, is submitted individually for synthetic generation and the synthetic results are appended to the initial synthetic dataset, generated from the original real data.

If unregulated, the first scenario opens up the possibility of making general inferences on the synthetic engine's behavior by investigating each new iteration of the synthetic data,

making comparisons across generations and comparing newly generated records which were part of a given update. From a use-case perspective, submitting the collective data for a full refresh (generating an entirely new dataset) will likely disrupt any ongoing analysis that was started on the original synthetic output. One option may be to consider processing a complete refresh every x months to limit the impact in this regard.

In the second scenario, the synthetic data can quickly become unreliable: as separate subsets of the synthetic data are trained on different input data, the continued representativeness of the real data is of utmost importance, and where differences occur between updates, separate synthetic updates may be overfit to particular behaviors, affecting the real dataset at large.

From a privacy perspective, Privacy Hub makes the following requirements:

1. Where the entire dataset is to be regenerated at each update, Subsalt must require that all older versions of the synthetic data are removed to ensure that statistical inferences cannot be made against a growing number of synthetic datasets generated from similar real datasets. Privacy Hub assumes that Subsalt has appropriate data use agreements with its clients in this regard.

2. Where only the additional records, making up the update, are synthesized, Privacy Hub requires that both the synthetic records generated from the update, and the fully updated synthetic output (consisting of the original synthetic output with all appended synthetic updates), are assessed using all tests as detailed in Sections 3.7 and 3.10.

3.7 Similarity Measures

This section discusses the measurement of overall similarity between the real dataset and the synthetically generated dataset. It is imperative that a successful synthetic engine generates data that is sufficiently similar to the real data such that relationships within it are preserved, while not being so similar that the real data may be re-engineered. Hence, there is a so-called 'Goldilocks' similarity where these antagonistic requirements are both

met.

3.7.1 Distance to Closest Record

A popular approach to the measurement of this similarity is the distance to closest record (DCR) – a suitable measure for a synthetic dataset, both in terms of fidelity to the underlying real dataset and the privacy of the resulting synthetic data. However, it is a nuanced point as how best to implement and interpret this metric with regard to privacy concerns, while balancing the need to maintain utility. For example, the application of a blanket minimum distance across an entire dataset can, in certain circumstances, result in an unnecessarily large reduction in utility whilst limiting the privacy risk ineffectively. An example of this would be where no records within the synthetic dataset are allowed to be within less than DCR = 2 to records within the real dataset. At the same time, a constant DCR can be an indicator that the same modification has been applied to some of the real variables during the generation of the synthetic version; an example of such a modification would be where the age variable is multiplied by a constant, thereby making it easier to back-engineer the real values.

Additionally, if the real dataset is small enough and the combinations of variable values are limited, the synthetic dataset will likely reproduce some of the records within the real dataset, resulting in a DCR of 0. While the presence of records with DCR = 0 translates into real records being replicated within the synthetic dataset, this does not implicitly mean that the privacy of those individuals recurrent in both the real and synthetic datasets has been compromised. For many such individuals within the real population, even though their information has been reproduced within the synthetic dataset, it may remain difficult to distinguish them from other individuals – e.g., where the real dataset only contains information about patient gender and race, which in combination with reasonably available information is very unlikely to re-identify an individual. Also, synthetic records which are similar to some underlying real records and, thus, have a 'low' DCR may not constitute a privacy breach if the corresponding real record is part of a large equivalence class of near-similar records, i.e., if the information gain from such an inference is low and related to information that falls towards the center of the distribution of values within the dataset. In contrast to this, upon calculation of DCR in some

outlier subgroups, and despite those synthetic records having a 'large enough' DCR, the information gain here can be valuable if, for instance, certain values which are close to those reproduced in the synthetic record allow the real individual(s) to be re-identified. For example, if one real record which presents an age value relating to a group of supercentenarians can be mapped to a (group of) specific synthetic record(s), even though the difference between the real and synthetic records' DCR appears to be large, the attacker could gain knowledge of previously unknown patient characteristics. Alternatively, as described in Section 3.8, there are cases in which a larger than one equivalence class within the synthetic dataset in combination with a DCR of 0 can constitute a privacy risk. Thus, the balance between DCR and equivalence class size must be carefully considered. This is discussed in more detail in Section 3.8.

In order to ensure that any records copied from the training dataset into the synthetic dataset present a sufficiently low risk of re-identification of the associated patients, Privacy Hub requires that there are less than 1% of records which i) have a DCR = 0, calculated using the Hamming distance, and ii) are in equivalence classes of five or less in the real data.

For those records with a DCR that is not equal to 0, additional considerations must be made to ensure that these records are sufficiently protected from re-identification. It is Privacy Hub's judgment that the DCR (between the real and synthetic records) should be compared to a meaningful benchmark, which in this case can be represented by the DCR of a fraction of the real data itself (outwith that of the training dataset). By this rationale, synthetic records should not be more 'similar' to the real training records than they are to a 'holdout' subset of the real data that was not used for training. Thus, a good test for the DCR measure between the real and synthetic datasets ($DCR(R, S)$) is to compare it to the DCR between the holdout and synthetic datasets ($DCR(H, S)$) where the two distributions are required to be close to identical. Moreover, the difference between the two resulting distributions of $DCR(R, S)$ and $DCR(H, S)$ can be tested using statistical hypothesis tests.

Firstly, however, an important consideration for any DCR metric is to decide which distance measure should be employed. In the context of health data, the selection of distance

metrics for calculating the DCR is influenced by the fact that this distance should be adjusted to deal optimally with both categorical and numeric data. As such, we only discuss measures able to accommodate for both categorical and numeric variable types.

Hamming Distance

The Hamming distance is used as a measure of the difference between categorical variables and those which may be mapped to categorical variables. For a given real record, r, and synthetic record, s, the values of each categorical variable, i, are compared. The Hamming distance, d_h, is then defined as the sum of individual variable distances, d_i, over all, n, variables:

$$d_h = \sum_{i=1}^{n} d_i \qquad \text{where} \qquad d_i = \begin{cases} 0 & \text{if } r_i = s_i \\ 1 & \text{if } r_i \neq s_i \end{cases}$$

In the above, we assume (without loss of generality) that r and s include only categorical variables. In reality, in order to use the Hamming distance as a DCR metric effectively, non-categorical variables must be converted to categorical variables where possible; otherwise, a combination of distance measures must be used.

Before applying the Hamming distance to numeric variables, the numeric values must be converted to categorical values. Without a carefully considered conversion, the effectiveness of the Hamming distance will be limited, particularly where the numeric range has an inherent meaning. For example, where the Hamming distance between two age variables is calculated, if each age is treated as a separate category, with $\text{age}_1 = 100$ and $\text{age}_2 = 10$, the resulting Hamming distance will be the same as that calculated between $\text{age}_1 = 100$ and $\text{age}_2 = 99$. In this example, it would make sense to define the categories more appropriately. Thus, in order to be able to include numeric variables within the DCR metric using the Hamming distance, the values populating numeric variables must be converted to meaningful categorical values. This may be achieved by using equal-depth bins for the values within the numeric variables. By splitting the values in such a manner, these bins may be treated as categorical values, allowing the Hamming distance to account adequately for both categorical and numeric variables. This is discussed in the sections to follow.

Gower Distance

The Gower distance is designed to be used to measure differences between a mix of categorical and numerical variables, as it is a combination of the Hamming distance (used for categorical and logical variables) and the Manhattan distance (used for numerical variables). The Gower distance is normalized to give a value between 0 and 1, and is defined as:

$$d_g = \sum_{i=1}^{n} \frac{d_i}{n} \qquad \text{where} \begin{cases} \text{if } i \text{ is categorical} & d_i = \begin{cases} 0 & \text{if } r_i = s_i \\ 1 & \text{if } r_i \neq s_i \end{cases} \\ \text{if } i \text{ is numeric} & d_i = \dfrac{|r_i - s_i|}{R_i} \end{cases}$$

where R_i is the range of the numerical variable i, and n is the number of total variables.

Privacy Hub considers that with an appropriate number of equal-depth bins, both the Hamming distance and the Gower distance are highly effective tools when used to calculate the distance between records where both categorical and numerical variables are part of the dataset. Moreover, the choice of distance metric can be influenced by scaling and variance within the variables. For example, the Gower distance is not prone to variation as it is not affected by user configuration (as is the case with the Hamming distance when the number of bins is chosen), while it is also normalized between a value of 0 and 1. Conversely, the Hamming distance is not normalized and can only take values of positive integers, making it easier to compare across different datasets.

3.7.2 Similarity Measures Testing

In this section, we present analysis of the behavior of the synthetic engine with regard to the similarity measures presented above. Privacy Hub has performed testing on a number of datasets, the results for five of which are detailed below.

The five datasets generated by Privacy Hub and used for testing contained both categorical and numerical variables, making them representative of most datasets which contain health information. As such, both the Gower and Hamming distances (with numerical variables either rounded or aggregated into bins) were the most appropriate measures for

making a determination on the similarity of the respective datasets. The results were reviewed by Privacy Hub and are detailed below.

3.7.3 Dataset 1

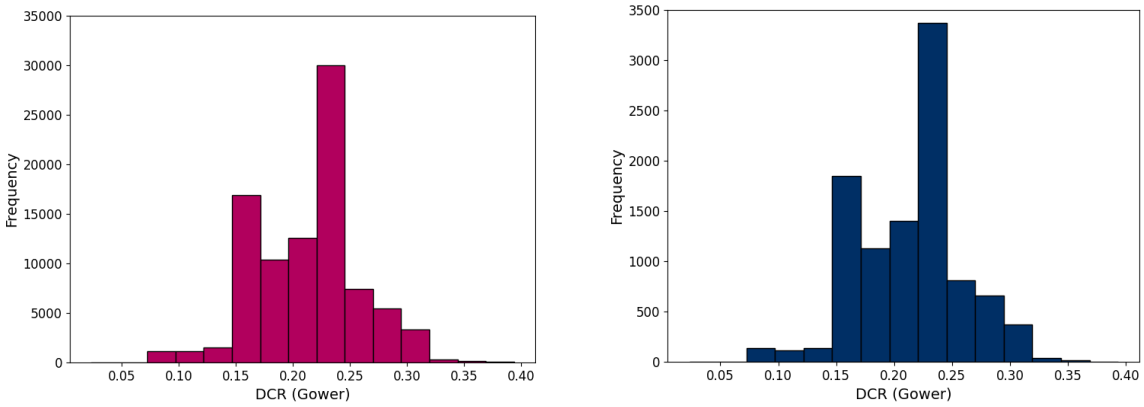
An initial test was conducted on the Privacy Hub-generated Dataset 1. The variables contained in this dataset are summarized below; the numeric variables are marked with a '*', while the remaining variables are categorical.

*age**
*year_of_birth**
marriage_status
gender
first_name
zip5
state
race
education
income
diagnosis_code
registered_voter
registered_car_owner
*height**
zip3

The dataset originally contained 100,000 rows, which were to be treated as individual patient records. However, Privacy Hub randomly split this original 'real' dataset into a holdout dataset containing 10,000 records and a training dataset containing 90,000 records. The holdout dataset was not shared with Subsalt. Once the holdout data was removed from the test dataset, the remaining 90,000 records that constituted the training dataset were sent to Subsalt. Following the delivery of this 'Training Dataset 1', Subsalt provided Privacy Hub with three synthetic datasets, one for each model (TVAE, CTGAN and CopulaGAN), each of which contained the variables detailed above and 90,000 rows.

The DCR between 'Training Dataset 1' and each of the three synthetic datasets, in addition to the DCR between 'Holdout Dataset 1' and the synthetic datasets, have been computed by Privacy Hub using the Gower distance. A summary of the results is shown in Table 1. Example plots for the CTGAN model are shown in Figures 1a and 1b, respectively. There were no identical records (DCR = 0) between the datasets (training, holdout and the three synthetic datasets). However, for the TVAE model, the minimum

Figure 1 – DCR Dataset 1 – Gower example



(a) Histogram of DCR between 'Training Dataset 1' and the synthetic dataset for the CTGAN model based on Gower distance

(b) Histogram of DCR between 'Holdout Dataset 1' and the synthetic dataset for the CTGAN model based on Gower distance

DCR between the training data and the synthetic data was 0.007, while the minimum DCR between the holdout and the synthetic data was 0.033. This indicates that some records were very close to being identical, with the very small value of the minimum DCR being based on differences between the values of the numerical variables. For all three models, the mean values of the two DCRs ($DCR(R, S)$ and $DCR(H, S)$) were found to be reasonably similar with differences of less than 0.003.

Table 1 – DCR Dataset 1 summary – Gower

		DCR				
Model	Training/Holdout	Mean	Max	Min	Std. dev.	DCR0 count
TVAE	Training	0.295	0.540	0.007	0.070	0
	Holdout	0.297	0.504	0.033	0.069	0
CTGAN	Training	0.213	0.394	0.023	0.045	0
	Holdout	0.214	0.366	0.034	0.045	0
CopulaGAN	Training	0.202	0.388	0.008	0.043	0
	Holdout	0.203	0.378	0.026	0.045	0

Following this, in order to test whether the distributions of the DCR measures (based on Gower distance between 'Training Dataset 1' and the synthetic datasets, and between

'Holdout Dataset 1' and the synthetic datasets) were significantly different, the Kolmogorov–Smirnov (KS) test was performed. Furthermore, a t-test to assess whether the two distributions (for each model) had the same mean was also performed. The results are shown in Table 3.

For all three models, the KS test provided results that support the null hypothesis that two samples come from the same distribution. However, in the t-test of both the TVAE and CTGAN models, the results do not support the null hypothesis that the difference between the two means is zero. The null hypothesis is supported for the CopulaGAN model.

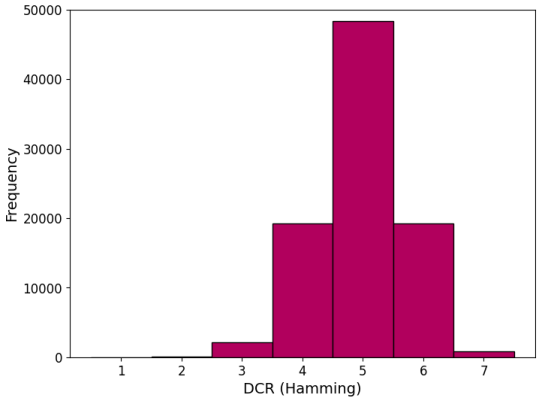
Next, the DCR was calculated in a similar manner as described above, this time using the Hamming distance. For testing, the *age* and *year_of_birth* variables were left unchanged, matching only on exact matches between real and synthetic datasets, whilst the *height* variable was recoded into 25 equal-depth bins.

Table 2 – DCR Dataset 1 summary – Hamming

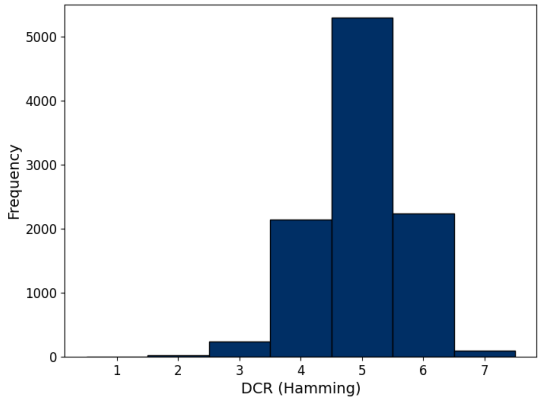
		DCR				
Model	Training/Holdout	Mean	Max	Min	Std. dev.	DCR0 count
TVAE	Training	5.799	9	1	1.059	0
	Holdout	5.811	9	2	1.042	0
CTGAN	Training	4.967	8	1	0.756	0
	Holdout	4.976	7	2	0.764	0
CopulaGAN	Training	4.840	8	1	0.723	0
	Holdout	4.839	7	1	0.727	0

The DCR between 'Training Dataset 1' and each of the three synthetic datasets, in addition to the DCR between 'Holdout Dataset 1' and the synthetic datasets, have been computed by Privacy Hub using the Hamming distance. A summary of the results is shown in Table 2. Example plots for the CTGAN model are shown in Figures 2a and 2b, respectively. There were no identical records (DCR = 0) between the datasets (training, holdout and the three synthetic datasets). For the TVAE and CTGAN models, the minimum (Hamming) DCR between the training and the synthetic data was 1, which implies that there were records where only one categorical value differed between the real and

Figure 2 – DCR Dataset 1 – Hamming example



(a) Histogram of DCR between 'Training Dataset 1' and the synthetic dataset for the CTGAN model based on Hamming distance



(b) Histogram of DCR between 'Holdout Dataset 1' and the synthetic dataset for the CTGAN model based on Hamming distance

synthetic records. For the same models, the minimum DCR between the holdout and the synthetic data was 2, meaning there were closer matches between records in the synthetic data and the training data than between the synthetic data and the holdout data. However, on average the two DCRs ($DCR(R,S)$ and $DCR(H,S)$), for all three models, were found to be reasonably similar. To explore this further, the t-test was performed (the Kolmogorov–Smirnov test is not a suitable measure for discrete data). The results show that for each model, the null hypothesis that the two samples have the same mean is supported. Again, on average, the synthetic datasets are as similar to 'Training Dataset 1' in terms of DCR as they are to 'Holdout Dataset 1'.

Table 3 – DCR statistics: Dataset 1

	T-test (Gower)		T-test (Hamming)		KS test (Gower)		
Model	Statistic	Precision	Statistic	Precision	Statistic	Precision	Pass/Fail (P/F)
TVAE	-2.843	0.004	-1.021	0.307	0.013	0.077	FPP
CTGAN	-2.140	0.032	-1.178	0.234	0.014	0.063	FPP
CopulaGAN	-1.029	0.304	0.087	0.930	0.008	0.560	PPP

Comment on 'subsalt_other' and Note on the Updated Synthetic Datasets

During the early part of this process, a major feature of the synthetic datasets generated

by Subsalt was the presence of 'subsalt_other' values. Indeed, large proportions of the early training data fell below the occurrence threshold and were therefore converted to 'subsalt_other' prior to synthetic generation as detailed in Section 3.2.13. In later testing, the threshold was lowered significantly, all but removing the presence of 'subsalt_other' values. In this scenario, a low proportion of 'subsalt_other' values was initially present, the results for which are detailed above. Subsequently, Privacy Hub removed the records containing 'subsalt_other' from all three synthetic datasets; the results were analyzed and found to be close to those presented above. The low proportion of 'subsalt_other' values had minimal impact on the DCR-related results. The proportion of records removed for each synthetic dataset was as follows: for the TVAE model, 118 records corresponding to 0.13%; for the CTGAN model, 160 records corresponding to 0.18%; for the CopulaGAN model, 11 records corresponding to 0.01%.

3.7.4 Dataset 2

The second test dataset was generated from Test Dataset 1, consisting of the same variables and 100,000 records which were again split into a training (90,000) and a holdout (10,000) dataset. Subsalt provided Privacy Hub with three synthetic datasets, one for each model (TVAE, CTGAN and CopulaGAN), each of which contained the variables detailed above and 90,000 rows. The training data was manufactured to contain added duplicate and similar groups. Table 4 gives the details of the added group sizes.

The DCR between 'Training Dataset 2' and each of the three synthetic datasets, in addition to the DCR between 'Holdout Dataset 2' and the synthetic datasets, have been computed by Privacy Hub using both the Gower and Hamming distances. A summary of the results is shown in Tables 5 and 6. Again, for testing, the *age* variable was left unchanged, matching only on exact matches between real and synthetic datasets, whilst the *height* variable was recoded into 25 equal-depth bins. Example plots for the CTGAN model using Gower and Hamming distances are shown in Figures 3 and 4, respectively. There were no identical records (DCR = 0) between the datasets (training, holdout and the three synthetic datasets) using the Gower calculation. However, the minimum DCR in each model was very small, indicating the presence of records that are essentially identical, likely differing only slightly in some numeric variable. Using the hamming DCR, there

Table 4 – DCR Dataset 2 summary – Gower

Group	Matching records	Group size
1	15	100
2	15	500
3	15	1,000
4	15	2,000
5	15	5,000
6	14	3,000
7	13	5,000
8	12	5,000
9	12	7,000
10	11	10,000
11	8	15,000

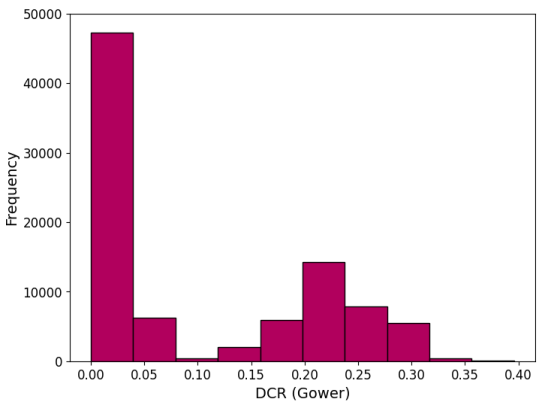
were greater than 20,000 identical records matching between the training and synthetic data, and yet no identical records between the holdout and synthetic datasets. Additionally, in both calculation scenarios, there was a large difference between the mean DCR for the training and holdout datasets, and also a large difference in minimum DCR between training and holdout wherein the minimum DCR between the synthetic and training datasets was much smaller than that of the holdout and synthetic datasets. These results clearly show that the synthetic data was overfit to the training data, resulting in both identical and very similar records appearing in the synthetic output.

Table 5 – DCR Dataset 2 summary – Gower

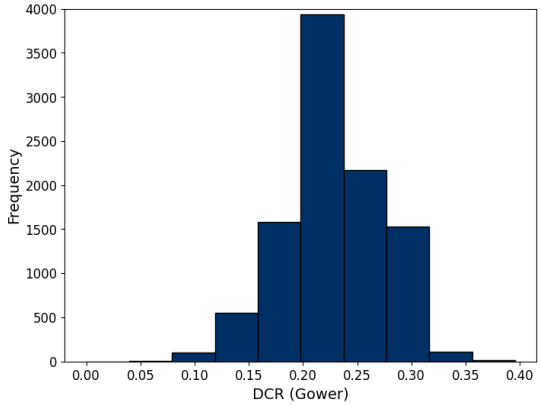
		DCR				
Model	Training/Holdout	Mean	Max	Min	Std. dev.	DCR0 count
TVAE	Training	0.107	0.458	2.609e-09	0.120	0
	Holdout	0.247	0.445	0.074	0.058	0
CTGAN	Training	0.099	0.386	2.086e-08	0.113	0
	Holdout	0.230	0.396	0.013	0.044	0
CopulaGAN	Training	0.099	0.431	6.369e-08	0.111	0
	Holdout	0.226	0.397	0.060	0.046	0

Once again, the Kolmogorov–Smirnov test was performed in order to test whether the distributions of the DCR measures (based on Gower distance between 'Training Dataset 2'

Figure 3 – DCR Dataset 2 – Gower example

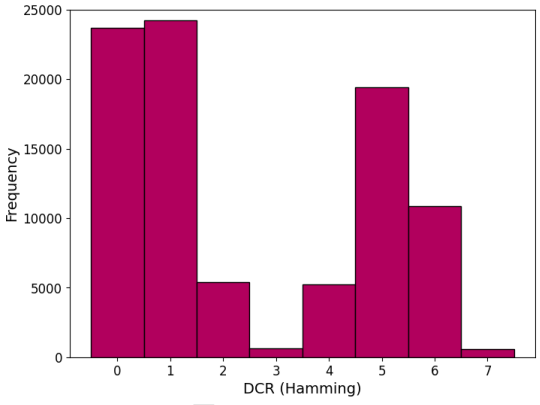


(a) Histogram of DCR between 'Training Dataset 2' and the synthetic dataset for the CTGAN model based on Gower distance

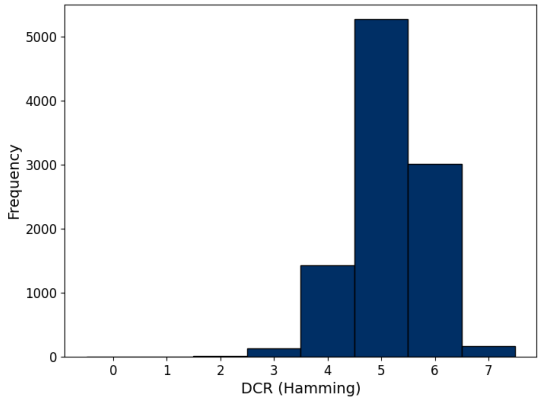


(b) Histogram of DCR between 'Holdout Dataset 2' and the synthetic dataset for the CTGAN model based on Gower distance

Figure 4 – DCR Dataset 2 - Hamming example



(a) Histogram of DCR between 'Training Dataset 2' and the synthetic dataset for the CTGAN model based on Hamming distance



(b) Histogram of DCR between 'Holdout Dataset 2' and the synthetic dataset for the CTGAN model based on Hamming distance

Table 6 – DCR Dataset 2 summary – Hamming

		DCR				
Model	Training/Holdout	Mean	Max	Min	Std. dev.	DCR0 count
TVAE	Training	2.408	8	0	2.390	27,548
	Holdout	5.216	8	2	0.879	0
CTGAN	Training	2.488	8	0	2.314	23,696
	Holdout	5.163	7	2	0.733	0
CopulaGAN	Training	2.473	8	0	2.308	23,955
	Holdout	5.121	8	2	0.762	0

and the synthetic datasets, and between 'Holdout Dataset 2' and the synthetic datasets) were significantly different. The results (shown in Table 7) across all three models show clearly that the null hypothesis is not supported and that the DCR results were not drawn from the same distribution. Furthermore, t-tests were performed for each model using both the Gower and Hamming DCR. Again, the results (also shown in Table 7) demonstrate that for each model, in each calculation scenario, the mean of the distributions was not equal.

Table 7 – DCR statistics: Dataset 2

	T-test (Gower)		T-test (Hamming)		KS test (Gower)		
Model	Statistic	Precision	Statistic	Precision	Statistic	Precision	Pass/Fail (P/F)
TVAE	-115.052	0.000	-116.600	0.000	0.595	0.000	FFF
CTGAN	-115.214	0.000	-114.988	0.000	0.594	0.000	FFF
CopulaGAN	-113.373	0.000	-114.025	0.000	0.587	0.000	FFF

3.7.5 Dataset 3

The next test was conducted on the Privacy Hub-generated Dataset 3. The dataset contained 20 variables, summarized below; the numeric variables are marked with a '*', while the remaining variables are categorical.

The dataset originally contained 100,000 rows, which were to be treated as individual patient records and which contained multivariate relationships to better represent a real dataset likely to be seen by Subsalt. Privacy Hub randomly split this original 'real' dataset

into a holdout dataset containing 50,000 records and a training dataset containing 50,000 records. The holdout dataset was not shared with Subsalt. Once the holdout data was removed from the test dataset, the remaining 50,000 records that constituted the training dataset were sent to Subsalt. Following the delivery of this 'Training Dataset 3', Subsalt provided Privacy Hub with three synthetic datasets, one for each model (TVAE, CTGAN and CopulaGAN), each of which contained the variables detailed above and 50,000 rows.

*age**
*height**
rating
race
gender
owner
response1
response2
marstat
customer
*yob**
diagnosis0
diagnosis1
diagnosis2
diagnosis3
diagnosis4
diagnosis5
zip3
providerzip3
income

The DCR between 'Training Dataset 3' and each of the three synthetic datasets, in addition to the DCR between 'Holdout Dataset 3' and the synthetic datasets, have been computed by Privacy Hub using the Gower distance. A summary of the results is shown in Table 8. Example plots for the CTGAN model are shown in Figures 5a and 5b, respectively. There were no identical records (DCR = 0) between the datasets (training, holdout and the three synthetic datasets), and indeed, all minimum DCR values were relatively high and closely matched with the results for the holdout dataset (with slight variation for the CopulaGAN model).

Once again, in order to test whether the distributions of the DCR measures (based on Gower distance between 'Training Dataset 3' and the synthetic datasets, and between 'Holdout Dataset 3' and the synthetic datasets) were significantly different, the Kolmogorov–Smirnov test was performed. Furthermore, a t-test to assess whether the two distributions (for each model) had the same mean was also performed. The results are

Table 8 – DCR Dataset 3 summary – Gower

		DCR				
Model	Training/Holdout	Mean	Max	Min	Std. dev.	DCR0 count
TVAE	Training	0.466	0.725	0.254	0.062	0
	Holdout	0.468	0.715	0.254	0.062	0
CTGAN	Training	0.437	0.578	0.305	0.031	0
	Holdout	0.438	0.604	0.309	0.031	0
CopulaGAN	Training	0.430	0.581	0.279	0.032	0
	Holdout	0.430	0.573	0.305	0.032	0

shown in Table 10.

For both the CTGAN and CopulaGAN models, the KS test provided results that support the null hypothesis that two samples come from the same distribution, and the t-test results support the hypothesis that the means were the same. However, for the TVAE model, the results do not support either hypothesis.

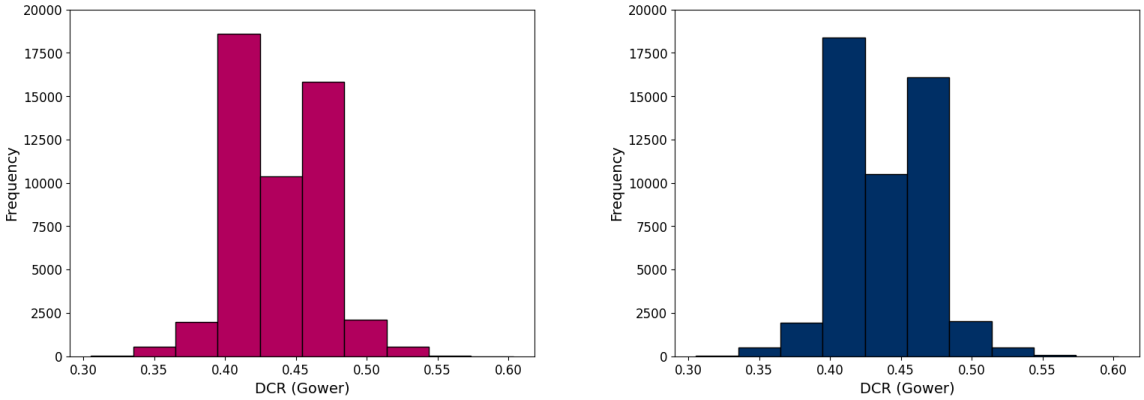
Next, the DCR was calculated in a similar manner as described above, this time using the Hamming distance. For testing, the *age* variable was left unchanged, matching only on exact matches between real and synthetic datasets, whilst the *height* variable was recoded into 25 equal-depth bins.

Table 9 – DCR Dataset 3 summary – Hamming

		DCR				
Model	Training/Holdout	Mean	Max	Min	Std. dev.	DCR0 count
TVAE	Training	11.302	16	6	1.267	0
	Holdout	11.321	16	7	1.254	0
CTGAN	Training	10.936	14	8	0.693	0
	Holdout	10.942	14	8	0.686	0
CopulaGAN	Training	10.817	14	8	0.720	0
	Holdout	10.820	14	8	0.712	0

The DCR between 'Training Dataset 3' and each of the three synthetic datasets, in addition to the DCR between 'Holdout Dataset 3' and the synthetic datasets, have been computed by Privacy Hub using the Hamming distance. A summary of the results is

Figure 5 – DCR Dataset 3 – Gower example



(a) Histogram of DCR between 'Training Dataset 3' and the synthetic dataset for the CTGAN model based on Gower distance

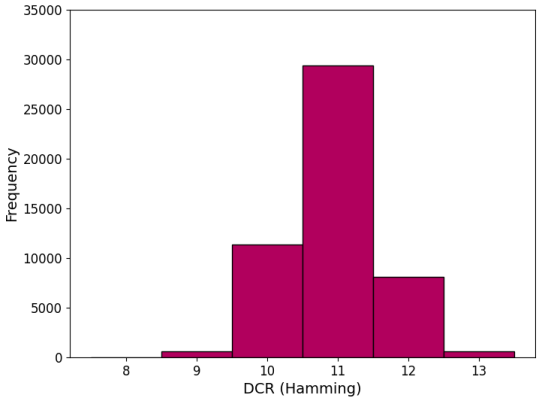
(b) Histogram of DCR between 'Holdout Dataset 3' and the synthetic dataset for the CTGAN model based on Gower distance

shown in Table 9. Example plots for the CTGAN model are shown in Figures 6a and 6b, respectively. There were no identical records (DCR = 0) between the datasets (training, holdout and the three synthetic datasets). For the TVAE model, the minimum DCR between the training and the synthetic data was 6, whereas the minimum DCR between the holdout and the synthetic data was 7; i.e., there were closer matches between records in the synthetic data and the training data than between the synthetic and holdout data. The t-test was performed; the results show that for the CTGAN and CopulaGAN models, the null hypothesis that the two samples have the same mean is supported. This is not the case for the TVAE model.

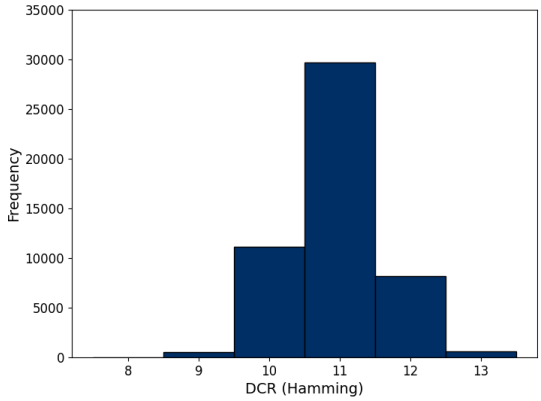
Table 10 – DCR statistics: Dataset 3

	T-test (Gower)		T-test (Hamming)		KS test (Gower)		
Model	Statistic	Precision	Statistic	Precision	Statistic	Precision	Pass/Fail (P/F)
TVAE	-2.774	0.006	-2.280	0.023	0.012	0.002	FFF
CTGAN	-0.859	0.390	-1.353	0.176	0.006	0.260	PPP
CopulaGAN	0.245	0.806	-0.548	0.584	0.004	0.894	PPP

Figure 6 – DCR Dataset 3 – Hamming example



(a) Histogram of DCR between 'Training Dataset 3' and the synthetic dataset for the CTGAN model based on Hamming distance



(b) Histogram of DCR between 'Holdout Dataset 3' and the synthetic dataset for the CTGAN model based on Hamming distance

3.7.6 Dataset 4 - Removal of Multivariate Relationships

A fourth test dataset was generated in a similar manner to Dataset 3, consisting of the same variables (treated in the same way for testing) and 100,000 records. In this testing scenario, multivariate relationships were not included. The data was split into a training (50,000) and a holdout (50,000) dataset as before. Subsalt provided Privacy Hub with three synthetic datasets, one for each model (TVAE, CTGAN and CopulaGAN), each of which contained the variables detailed above and 50,000 rows.

The results (Table 11) were not significantly different from the results generated on Dataset 3 (Table 10). However, the TVAE model (which failed two tests described above) was shown to pass all three tests, demonstrating that multivariate relationships do indeed have an effect on the outcome, and that the tests capture changes associated with such relationships.

Table 11 – DCR statistics: Dataset 4

	T-test (Gower)		T-test (Hamming)		KS test (Gower)		
Model	Statistic	Precision	Statistic	Precision	Statistic	Precision	Pass/Fail (P/F)
TVAE	-2.690	0.007	-1.688	0.091	0.008	0.086	FPP
CTGAN	-0.927	0.354	-0.619	0.536	0.005	0.579	PPP
CopulaGAN	-1.471	0.141	-1.887	0.059	0.008	0.119	PPP

3.7.7 Dataset 5

The final test was conducted on the Privacy Hub-generated Dataset 5. This dataset was also used to test membership inference; see Section 3.9.2 for more details[2]. The variables contained in this dataset are summarized below; the numeric variables are marked with a '*', while the remaining variables are categorical.

*Age**
ZIP3
Gender
*Income**
*Height**
Race
Marstat[3]
LOINC
Service_date
Allergies
*Days_in_hospital**
Facility_id

The dataset originally contained 100 rows which were to be treated as individual patient records. Privacy Hub randomly split this original 'real' dataset into a 'Holdout Dataset 5' which represented a random subset of the original dataset (containing the variables detailed above and 50 records), and 'Training Dataset 5' which contained the remaining 50 records. The holdout dataset was not shared with Subsalt. Following the removal of the holdout dataset, the remaining 50 records ('Training Dataset 5') were delivered to Subsalt, who subsequently provided Privacy Hub with three synthetic datasets, one for each model (TVAE, CTGAN and CopulaGAN), each of which contained the same variables and 50 rows.

[2]These results were calculated on the updated synthetic dataset in which the data class of the *ZIP3* and *Facility_id* variables matched.

[3]Contained values regarding patient marital status.

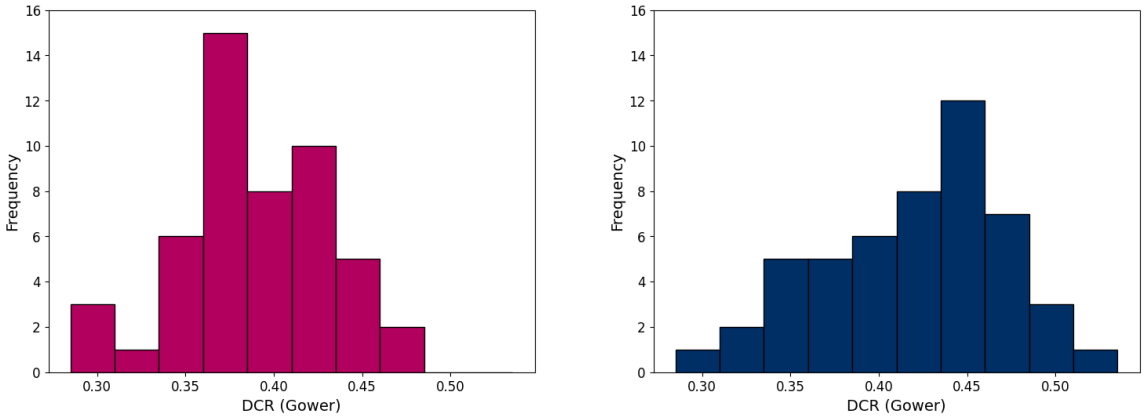
The DCR between 'Training Dataset 5' and each of the three synthetic datasets, in addition to the DCR between 'Holdout Dataset 5' and the synthetic datasets, have been computed by Privacy Hub using both the Gower and Hamming distances. A summary of the results is shown in Tables 12 and 13, respectively. Example plots for the CTGAN model using Gower and Hamming distances are shown in Figures 7 and 8, respectively. There were no identical records (DCR = 0) between the datasets (training, holdout and the three synthetic datasets) in either calculation scenario. However, as with Dataset 2, the results show that the engine was overfit to the training dataset, wherein the results are generally skewed towards smaller DCR values between the training and synthetic data than between the holdout and synthetic data. The mean DCR values are smaller, as are the minimum DCR values. As described in Section 3.9.2, this dataset was designed to fail a membership inference test. However, as DCR underpins both calculations, it is not surprising to see these results. We would expect that in such a small dataset, consisting of only 50 records, all models will skew towards overfitting to the training data. The results are not expected to be as extreme as described in Section 3.7.4, given the absence of duplicate records. Indeed, all records in this dataset may in a way be thought of as outlier records, given the size of the dataset and absence of trends therein.

Table 12 – DCR Dataset 5 summary – Gower

		DCR				
Model	Training/Holdout	Mean	Max	Min	Std. dev.	DCR0 count
TVAE	Training	0.316	0.531	0.150	0.088	0
	Holdout	0.404	0.565	0.267	0.080	0
CTGAN	Training	0.391	0.470	0.298	0.042	0
	Holdout	0.420	0.535	0.285	0.052	0
CopulaGAN	Training	0.396	0.547	0.223	0.072	0
	Holdout	0.418	0.556	0.265	0.067	0

As expected, the results of both the Kolmogorov–Smirnov and t-tests show this overfitting, with the only passing scenarios occurring for the CopulaGAN model in regard to the Gower t-test and the KS test. No other scenarios provide enough evidence to support the respective null hypotheses – meaning that, in general, the DCR results were not drawn from the same distribution and nor did they have the same mean.

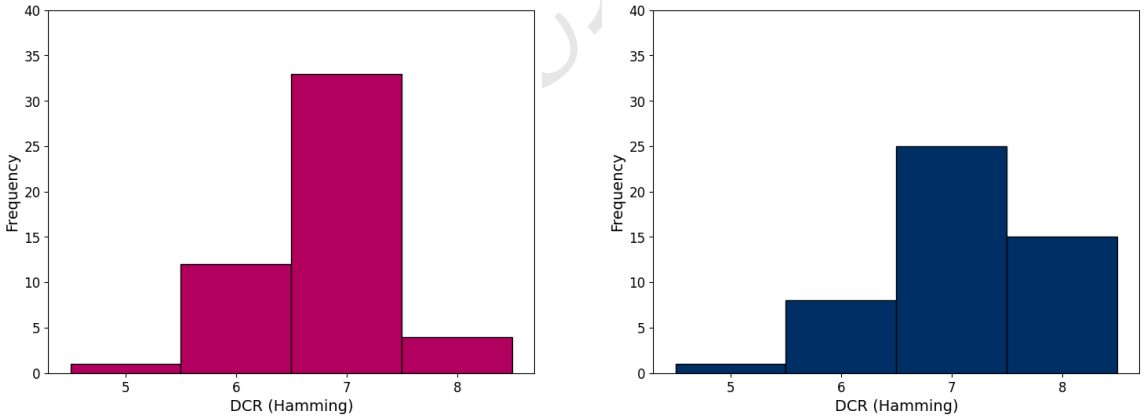
Figure 7 – DCR Dataset 5 – Gower example



(a) Histogram of DCR between 'Training Dataset 5' and the synthetic dataset for the CTGAN model based on Gower distance

(b) Histogram of DCR between 'Holdout Dataset 5' and the synthetic dataset for the CTGAN model based on Gower distance

Figure 8 – DCR Dataset 5 – Hamming example



(a) Histogram of DCR between 'Training Dataset 5' and the synthetic dataset for the CTGAN model based on Hamming distance

(b) Histogram of DCR between 'Holdout Dataset 5' and the synthetic dataset for the CTGAN model based on Hamming distance

Table 13 – DCR Dataset 5 summary – Hamming

		DCR				
Model	Training/Holdout	Mean	Max	Min	Std. dev.	DCR0 Count
TVAE	Training	6.140	9	4	1.125	0
	Holdout	6.920	9	5	1.007	0
CTGAN	Training	6.800	8	5	0.606	0
	Holdout	7.140	9	5	0.783	0
CopulaGAN	Training	6.640	8	5	0.898	0
	Holdout	7.000	8	5	0.904	0

Table 14 – DCR statistics: Dataset 5

	T-test (Gower)		T-test (Hamming)		KS test (Gower)		
Model	Statistic	Precision	Statistic	Precision	Statistic	Precision	Pass/Fail (P/F)
TVAE	-5.219	1.005e-06	-3.653	0.000	0.440	9.909e-05	FFF
CTGAN	-3.070	0.003	-2.429	0.017	0.340	0.006	FFF
CopulaGAN	-1.561	0.122	-1.998	0.048	0.220	0.179	PFP

3.7.8 Quantification of Disclosure Risk When DCR = 0

In this section, we look again at the distance to closest record. A DCR of 0 means that a real record was reproduced in the synthetic output and, thus, the private information of a patient might have been disclosed as it is part of the synthetic dataset. However, depending on the properties of the real dataset, the presence of a number of reproduced real records within a synthetic output is occasionally unavoidable. This calls for further mitigation of disclosure risk.

We consider the reproduced records within the synthetic dataset to be a subset of the real dataset and, therefore, require that these records are HIPAA compliant. In particular, the disclosure risk of the reproduced records must be 'very small' as determined by HIPAA. Privacy Hub uses a threshold of less than 1% of the records being 'high risk', in order to determine whether or not a dataset presents 'very small' disclosure risk. A high-risk record is one which identifies an individual either uniquely or in a group with fewer than five other individuals. Privacy Hub, therefore, requires that all synthetic records with a DCR of 0 and which are within an equivalence class of less than five in the real data

are considered 'high-risk' records, and that the percentage of such 'high-risk' records determined in this way should be below the 1% threshold. Of course, due care should be taken, particularly for those datasets which come close to but do not exceed the 1% threshold. Although this does not constitute a disclosure breach, it is but one form of information gain from the synthetic output. Used in combination with other techniques, some of which are described in this report, it may be possible to infer greater information than was initially thought.

3.7.9 Requirements and Recommendations

Following analyses of the tests described in the previous sections, Privacy Hub requires and recommends the following modifications with regard to testing the similarity between any real dataset and the associated synthetically generated dataset. We require that a dataset should only be considered sufficiently low risk from a DCR perspective if:

- The number of records within the real dataset that have a DCR = 0 in relation to records in the synthetic dataset, and that are in an equivalence class of five or less than five, should be lower than 1% of the total number of records within the real dataset.
- The distribution of the DCR between the real (training) dataset and the synthetic dataset should not be significantly different to that between a holdout dataset (a randomly selected subset of the real dataset, not used for training) and the synthetic dataset.
- The real (training) dataset has a minimum size of 3,000 records, and a minimum threshold of 5% of records are used as a holdout dataset which can be used for testing as detailed herein.

Additionally, Privacy Hub makes the following recommendations:

- When calculating the DCR using the Hamming distance, numerical variables should be binned based on their relevance and sensitivity rather than applying a blanket '25-bins' approach. Privacy Hub recommends that for indirect identifier variables

such as age or year of birth, exact year values are used, while for other variables such as height, weight, income etc., between ten and 25 bins are used depending on the spread of the variable.

- Where the size of the real dataset allows, and where utility will not be unduly affected, as much as 50% of records should be used to form a holdout dataset as described herein. The proportion of records included in a holdout dataset may vary depending on the size of the dataset. Privacy Hub recommends that a minimum of 10% of records are used, except in scenarios where the real data is very small (of the order of a few thousand records). We allow discretion in this choice, setting a required lower threshold of 5% of records.

- Where the size of the real dataset is small or very small, and where the holdout dataset represents less than 50% of the total records, Privacy Hub recommends that the training dataset, R, should contain the same number of records as the holdout dataset, H, and should be randomly selected from the remaining records of the real dataset following the extraction of the holdout data. Testing of the two resulting DCR distributions ($DCR(R, S)$ and $DCR(H, S)$) should be repeated multiple times, with different records included within the training dataset, R, to ensure that the results are consistent and free of sampling bias.

3.8 Equivalence Classes

The analysis of equivalence class sizes in a synthetically generated output can provide useful insights from both a privacy and a utility perspective. Consideration must be given as to whether inferences can be made on the underlying real records due to the presence of groups of duplicate or similar records in the synthetic data, and whether the engine itself was the cause of the appearance of said groups. Additionally, in the absence of clear regulatory advice or an evolved industry consensus on disclosure risk in the context of synthetic data, analyses which can act as a reference point for the measure of disclosure risk evaluated under a HIPAA Expert Determination can be beneficial.

The size of equivalence classes in any synthetic dataset is inherently linked to the size of

equivalence classes within the real dataset as the goal of the synthetic data is to preserve the distributions of the real data; therefore, any necessary precautions or modifications may need to be applied pre- and/or post-processing as the creation of duplicate records may not be an artifact of the synthetic engine itself. Indeed, it is difficult to determine whether the reproduction of records within larger equivalence classes is the product of the synthetic engine being fidelitous to the real data, or is the result of the synthetic engine producing these records disproportionately. If it is the latter, and the synthetic engine is seen to inflate the equivalence class size, an attacker may be able to make inferences on a smaller real group, thus increasing the associated disclosure risk of the real records. It is, of course, reasonable to assume that a poorly constructed real dataset, with a small variance in values, will produce larger equivalence classes in any synthetically generated output. It is, therefore, important to determine whether the engine is simply echoing the structure of the underlying data or whether it is disproportionately reproducing such equivalence classes. However, regardless of whether the engine itself is at fault or whether it is the underlying data, action may be required to ensure the risk of re-identification remains sufficiently low.

Additionally, the question of whether or not the inference made by an attacker that a patient, or a group of identifiers, appears in the real data constitutes a privacy risk is also dependent on which equivalence classes are present within the synthetic data, and whether the attacker can gain meaningful information (see Section 3.10 for more information). If, for example, the larger equivalence classes within the synthetic dataset correspond to a patient or group of patients for whom attributes are well represented in the underlying data, then there may be little meaningful information gain. However, if the larger equivalence classes in the synthetic data correspond to a group of duplicate records in the real data, then it is possible that a single patient may be uniquely defined and readily identifiable (see Section 3.2.3). In the following analysis, we measure the reproduction rate of duplicate and similar records between real and synthetic data, measuring the group sizes in the synthetic output and assessing whether they carry additional disclosure risk.

3.8.1 Test Dataset 1

Privacy Hub sent two datasets to Subsalt, comprising the 14 variables shown below and 90,000 rows.

age	*education*
year_of_birth	*income*
marriage_status	*diagnosis_code*
gender	*registered_voter*
zip5	*registered_car_owner*
state	*height*
race	*zip3*

One of the datasets, referred to herein as the 'experiment' dataset, contained a number of manufactured duplicate or 'similar' groups of varying sizes, matching across differing proportions of variables. The purpose of inserting these groups into the experiment real dataset was to investigate their effect on synthetic generation: do we see such groups appearing in the synthetic output more frequently? Can an attacker make inferences on the structure of the real dataset based on the synthetic output? Details of the manufactured groups are shown below.

- Group 1 – Features matching: all; no. of rows: 100
- Group 2 – Features matching: all; no. of rows: 500
- Group 3 – Features matching: all; no. of rows: 1,000
- Group 4 – Features matching: all; no. of rows: 2,000
- Group 5 – Features matching: all; no. of rows: 5,000
- Group 6 – Features matching: 13; no. of rows: 3,000
- Group 7 – Features matching: 12; no. of rows: 5,000
- Group 8 – Features matching: 11; no. of rows: 5,000
- Group 9 – Features matching: 11; no. of rows: 7,000
- Group 10 – Features matching: 10; no. of rows: 10,000
- Group 11 – Features matching: 7; no. of rows: 15,000

The second dataset acted as a 'control' dataset and did not include any manufactured

groups, consisting instead of 90,000 records of which only three were duplicates. From the synthetic output generated from the control dataset we looked to benchmark the effect of the manufactured groups.

For each of the control and experiment datasets, three synthetic datasets were returned, with each output generated using a different model: CopulaGAN, CTGAN, and TVAE. All of the synthetic outputs contained 90,000 rows and the same columns as the real data.

The following naming conventions are used herein for the relevant datasets:

Table 15 – Naming conventions used for real and synthetic datasets

Dataset	Reference name
Real control dataset	c_real
Real experiment dataset	e_real
Control dataset generated by modelX	c_modelX
Experiment dataset generated by modelX	e_modelX

Duplicates

We looked first at how duplicate records (those matching on all features) were processed by the engine. In the real control data there were three duplicate records, each appearing in a group size of two. For each of the CTGAN and CopulaGAN generated datasets, there were no duplicate records appearing within the synthetic outputs. For the data generated using the TVAE model, there were 2,743 duplicate records appearing in the synthetic output. The largest group size across these duplicate records was seven. The mean group size for the dataset was 1.036.

It was important to determine whether these 2,743 records were copied directly from the real dataset, represented a similar group appearing in the real dataset, or represented a disproportionate generation by the model. None of the 2,743 duplicate synthetic records appeared in the real data; nor indeed were they similar enough to any of the three duplicate real records to indicate that the synthetic duplicates were generated from the real duplicate records. Whilst there appeared to be some degree of similarity in a number of these duplicate synthetic groups whereby the groups matched on the feature-value pairs shown

in Table 16, no records in the real data exactly matched these values, indicating that some group in the real data was disproportionately represented in the synthetic output.

Table 16 – Duplicate records – control data

Variable	Value
marriage_status	1
gender	2
zip5	35767
state	AL
race	White
education	other higher educational institution
diagnosis_code	A65
registered_voter	0
registered_car_owner	0

How do these results compare to the experiment data? The real experiment dataset contained 11 manufactured groups, matching on varying proportions of both variables and rows. Details of the matching records are shown in Tables 17 and 18, and in the subsequent analysis we consider their effect on the synthetic output and the implications for the privacy of the underlying real patient records.

The largest group of duplicate records contained 5,000 rows (5%). We first considered whether any of these records from the duplicate groups (1–4) appeared within the synthetic data. The duplicate counts for Groups 1–5 are summarized in Table 19.

The e_TVAE dataset contained 4,214 duplicate records, with the largest of these duplicate groups containing 279 records. Again it is important to understand the structure of these duplicate groups. Some 578 of the 4,214 duplicates also appeared in e_real, leaving a total of 3,636 purely 'synthetic' duplicates. Of those copied from the real data, 279 records matched Group 5, 24 matched Group 4, and one record matched Group 3. It is of course expected that the manufactured groups appear within the synthetic output. However, it was also found that 129 unique real records were duplicated in the synthetic data, with

Table 17 – Group features – duplicate records

Features	Group 1	Group 2	Group 3	Group 4	Group 5
age	48	10	67	66	0
year_of_birth	1974	2012	1955	1956	2022
marriage_status	2	0	1	1	0
gender	1	2	2	2	2
zip5	94237	72469	94574	92059	92135
state	CA	AR	CA	CA	CA
race	White	Other	Asian	Hispanic	Other
education	bachelors degree	other higher educational institution	bachelors degree	other higher educational institution	bachelors degree
income	$55,000	$100,000	$100,000	$50,000	$50,000
diagnosis_code	A303	A667	A241	A203	A5144
registered_voter	1	1	0	0	0
registered_car_owner	0	0	0	0	1
height	181	174	154	193	213
zip3	942	724	945	920	921
GROUP SIZE	100	500	1,000	2,000	5,000

Table 18 – Group features – similar records

Features	Group 6	Group 7	Group 8	Group 9	Group 10	Group 11
age	27	X	X	85	41	X
year_of_birth	1995	X	X	1937	1981	X
marriage_status	2	1	X	2	X	X
gender	2	2	2	1	X	X
zip5	85372	35504	91899	35161	36904	35295
state	AZ	AL	CA	AL	AL	AL
race	Other	Other	Native American	Hawaiian	X	Other
education	bachelors degree	other higher educational institution	bachelors degree	boarding school	X	bachelors degree
income	$60,000	$60,000	$70,000	X	$55,000	$85,000
diagnosis_code	A392	A232	A5442	X	A666	A031
registered_voter	0	1	0	X	0	X
registered_car_owner	0	0	1	0	0	X
height	X	142	159	189	209	X
zip3	853	355	918	351	369	352
GROUP SIZE	3,000	5,000	5,000	7,000	10,000	15,000

group sizes ranging from two to 31.

The e_CTGAN dataset contained 2,316 duplicate records, with the largest of these du-

Table 19 – Duplicate record counts for Groups 1–5

		Model		
Group	Real	e_TVAE	e_CTGAN	e_CopulaGAN
1	100	0	0	0
2	500	0	0	0
3	1,000	1	0	0
4	2,000	24	0	0
5	5,000	279	538	85

plicate groups containing 538 records. Here, 257 of the 2,316 duplicates also appeared in e_real, leaving a total of 2,059 purely 'synthetic' duplicates. Of those copied from the real data, 538 records matched Group 5, whilst there were no matches for Groups 1–4. It was also found that 15 unique real records were duplicated in the synthetic data, with group sizes ranging from two to five.

The e_CopulaGAN dataset contained 3,564 duplicate records, with the largest of these duplicate groups containing 113 records. In this case, 307 of the 3,564 duplicates also appeared in e_real, leaving a total of 3,257 purely 'synthetic' duplicates. Of those copied from the real data, 85 records matched Group 5, whilst there were no matches for Groups 1–4. It was also found that 81 unique real records were duplicated in the synthetic data, with group sizes ranging from two to six. Table 20 shows the breakdown of duplicate groups appearing across all datasets, showing the mean and maximum group sizes and a breakdown of the number of groups greater than a range of values.

The presence of duplicate records in the synthetic data generated from a *unique* real record represents a non-trivial risk of re-identification, and Privacy Hub requires that duplicate synthetic records generated from a unique real record are removed from the synthetic data during a post-processing phase. However, in datasets containing a smaller number of features, due to the resulting small number of possible combinations across the values of those features, it is possible that duplicate records appear which genuinely correspond to different patients, separated, for example, by a unique patient ID which is not carried through to the synthetic version. In such cases, provided the patient exists in

Table 20 – Duplicate counts

Dataset	Mean	Max	>1	>2	>5	>10	>50	>100	>500
Real control	1.00033	2	3	0	0	0	0	0	0
c_TVAE	1.0358	7	2,743	303	2	0	0	0	0
c_CTGAN	1.0000	1	0	0	0	0	0	0	0
c_CopulaGAN	1.0000	1	0	0	0	0	0	0	0
Real experiment	1.5759	5,000	2,582	1,013	641	538	155	35	3
e_TVAE	1.1862	279	4,214	1,627	525	272	33	2	0
e_CTGAN	1.1571	538	2,316	1,212	527	263	32	5	1
e_CopulaGAN	1.1495	113	3,564	1,400	530	243	10	1	0

a group size of five or greater within the real data (excluding the unique identifier), these duplicate records may be retained within the synthetic version. Indeed, where large group sizes exist within the synthetic data and are otherwise outlier groups (with regard to their group size), it should be determined whether, within the real data, the group pertains to a single patient or small groups of patients matching on a number of variables, or to a large group of distinct patients. Privacy Hub requires that the unique patients which form that group exist within a group size (as determined by the indirect identifiers) larger than five within the real data. Otherwise, the records must be removed from the synthetic data. Information on records with a DCR equal to zero can be found in Section 3.7.8.

Of course, the presence of manufactured groups such as those described herein will unduly influence the synthetically generated data. Creating large groups that match across eight features will affect the sizes of matching groups across all subsets of those features, which were generated from the same subgroups in the real data, while allowing for variation in the synthetically produced version. Indeed, in the e_TVAE example above, the group of synthetic duplicates within the group of size 31 matches with Group 9.

The presence of large duplicate groups in the real data has direct implications for the privacy of the underlying patients, whose data is at greater risk of reproduction by the synthetic engine. Privacy Hub requires that all duplicate records are removed from the real dataset during pre-processing, *before* generation of the synthetic data.

Similar Groups

In a similar manner, we can investigate how the manufactured 'similar' groups are represented in the synthetic output. It is important to understand, for example, what might be the effect of a real dataset containing 2% of records matching on 70% of features in the synthetic output, compared to a real dataset containing 5% of records matching on 30% of features. In the real dataset, Groups 6–11 consisted of varying numbers of partially matching records. For example, Group 8 contained 5,000 records matching on 11 variables. Table 21 shows the counts for these similar groups in both the real and synthetic datasets. It is clear that the synthetic data was not recreating the manufactured groups in numbers comparable to their appearance in the real data. In all instances, the group size of the manufactured groups was much lower in the synthetic output than in the real data. Group 11 was proportionally the closest representation, at least for the TVAE and CTGAN models. Group 8 was an outlier in the opposite direction, with distinctly fewer reproductions of these groups.

From a privacy perspective, that large groups of similar records are not reproduced at a similar proportion only helps to minimize the risk of re-identification arising from these records. This is perhaps an issue for utility, though we emphasise here that this experiment real dataset was manufactured to stress-test these effects and was not expected to be representative of a real-life dataset. However, an additional source of risk is that those records which occur within these larger group sizes are at a greater risk of a membership inference attack. More information about the risk associated with such attacks can be found in Section 3.9.

Table 21 – Record counts for Groups 6–11

		Model			
Group	No. matching features	Real	e_TVAE	e_CTGAN	e_CopulaGAN
6	13	3,000	710	15	1
7	12	5,000	1,192	1,658	565
8	11	5,000	209	59	198
9	11	7,000	1,536	572	724
10	10	10,000	2,065	2,855	1,297
11	7	15,000	11,357	4,363	12,588

Generic Groups

We extended this analysis to consider how these numbers looked, relative to other generic groups. In contrast to the previous section where we had explicit groupings within our datasets, in this analysis we compared generic group sizes matching across features rather than a particular set of values. For example, for Group 7, in the previous section we analyzed those records matching on a particular set of values for the defined features. Here, in contrast, we analyzed all matching records across the defined features, and considered all matching combinations of values rather than a specifically defined example.

Tables 22 and 23 show the group counts for generic matches on features defined by Groups 7 and 8. We can see clearly that in the control data the TVAE model was an outlier, with a much larger mean group size than the control data. For the experiment real data, all models displayed a smaller mean than the real data. Again, the TVAE model was somewhat of an outlier, producing a mean group size much closer to that of the real data than the other models.

Table 22 – Group counts for records matching Group 7

Dataset	Mean	Max	>1	>2	>5	>10	>50	>100	>500
Real control	1.0001	2	9	0	0	0	0	0	0
c_TVAE	1.2784	33	10,423	3,886	692	108	0	0	0
c_CTGAN	1.0005	2	45	0	0	0	0	0	0
c_CopulaGAN	1.0000	1	0	0	0	0	0	0	0
Real experiment	1.9136	5,000	4,299	2,334	1,011	522	94	39	7
e_TVAE	1.6119	1,192	7,638	4,346	1,690	678	57	9	1
e_CTGAN	1.2873	1,658	3,261	1,601	584	288	62	17	4
e_CopulaGAN	1.4099	565	4,895	2,911	1,363	624	49	8	1

Means

Finally, we looked more generally at the mean group size across all generated datasets, as shown in Table 24.

In general, the TVAE model was somewhat of an outlier, producing larger groups on

Table 23 – Group counts for records matching Group 8

Dataset	Mean	Max	>1	>2	>5	>10	>50	>100	>500
Real control	1.0001	2	9	0	0	0	0	0	0
c_TVAE	1.2785	33	10,424	3,886	692	108	0	0	0
c_CTGAN	1.0008	3	71	1	0	0	0	0	0
c_CopulaGAN	1.0000	1	0	0	0	0	0	0	0
Real experiment	2.0154	5,000	3,292	1,820	962	560	67	53	5
e_TVAE	1.7503	1,295	6,678	3,967	1,801	849	77	32	1
e_CTGAN	1.3282	1,901	3,066	1,549	643	285	77	29	4
e_CopulaGAN	1.4838	627	4,074	2,346	1,226	715	84	20	2

Table 24 – Mean group counts across all models

	Control				Experiment			
Group	Real	TVAE	CTGAN	CopulaGAN	Real	TVAE	CTGAN	CopulaGAN
1–5	1.0000	1.0358	1.0000	1.0000	1.5759	1.1862	1.1571	1.1495
6	1.0010	1.2947	1.0007	1.0000	1.9475	1.6723	1.2563	1.4344
7	1.0001	1.2785	1.0005	1.0000	1.9136	1.6119	1.2873	1.4099
8	1.0001	1.2785	1.0008	1.0000	2.0155	1.7503	1.3282	1.4838
9	1.0003	1.8245	1.0031	1.0007	1.7649	1.3294	1.2940	1.3123
10	1.0000	1.0582	1.0004	1.0000	1.6428	1.3199	1.2344	1.2682
11	1.0031	4.1599	1.0808	1.0048	2.2310	3.0960	1.7018	1.9704

average than the other models. The other outlier was Group 11 which, again, was generally produced in higher numbers relative to the other groups, this time across all models. Of course, Group 11 matched on the fewest features, leaving the highest proportion of 'free' variables to vary. We can expect that the equivalence classes of records generated from this group will depend on the depth of possible values in the free variables. For example, if the remaining free variables were all binary, we would expect a larger proportion of duplicates than if each free variable were a categorical variable of depth 500. In general, as the number of matching variables decreases, the further the risk of directly inferring information from the group size itself is reduced. It is of course possible to infer additional information about the record based on a subset of matching features; more information on the disclosure risk in this context can be found in Section 3.10.

Otherwise, for the control data, the mean group size was relatively close to the real data. For the experiment data, the mean group size was smaller than the real data. Again, it is not expected that this test dataset would be representative of real data.

3.8.2 Requirements and Recommendations

Following analyses of the tests described in the previous sections, Privacy Hub requires and recommends the following modifications with regard to the equivalence class sizes found within the synthetic dataset:

- Duplicate synthetic records generated from a unique real record must be removed from the synthetic data during a post-processing phase. A single occurrence of the record may be maintained in the synthetic data only if the record exists within a group size, as defined by the indirect identifiers, larger than five within the real data.
- For large equivalence classes, which are outlier groups with regard to their group size in the synthetic data, the unique patients (as defined by the indirect identifiers) forming the group must exist within an equivalence class larger than five within the real data. Otherwise, the records must be removed from the synthetic data.
- All duplicate records must be removed from the real dataset during pre-processing, *before* generation of the synthetic data.

Additionally, Privacy Hub makes the following recommendation:

- Privacy Hub recommends that outlier groups are defined as groups with a group size greater than three standard deviations from the mean value.

3.9 Membership Inference

Membership inference is a type of attack on synthetic data in which an attacker attempts to infer information about the real training dataset. In particular, the attacker seeks to determine, through analysis of the synthetic data, whether a real patient record known to

the attacker was part of the real population on which the synthetic engine was trained. This type of attack carries a particularly high degree of disclosure risk, especially if that risk takes into account the sensitivity of the health information involved, in cases where the training population is defined by specific characteristics; for example, where all patients have been diagnosed with cancer. In such cases, where it is revealed that a patient is a member of the training dataset, sensitive information relating to the patient is also disclosed.

In order to assess the risk associated with a membership inference attack, a test is constructed which aims to determine whether or not a random record was included within the training dataset. This test can be performed by splitting the 'real' dataset into two random and equal subsets, R and H, where R is the subset on which the synthetic model is trained, and H is a holdout dataset which is unknown to the synthetic engine. A synthetic dataset, S, is then generated by the engine, based on R. The assumption is that the attacker has access to a dataset RH (where the equal-sized subsets are amalgamated). In order to determine whether a record of the RH dataset was part of the R subset (and, thus, to infer its membership of the training dataset), distance metrics with appropriate thresholds can be used (other classification methods may be applied instead). Using these metrics, it is possible to determine if a synthetic record shares enough similarities with a given record, i (from the RH dataset), to identify that record as part of training dataset R. Explicitly, each record i of the RH dataset is labeled according to membership of the training subset R ('yes' or 'no'), based on the results of the application of a distance metric with an appropriate threshold.

For this type of attack, sensitivity is markedly biased towards detecting training dataset membership rather than non-membership. Thus, the classification problem's precision (ratio of true positives[4] to all positives for membership inference of R) is a useful measurement of the success of this attack. In assessing the membership inference risk associated with the correct inference that a record i is part of R, a precision closer to 1 indicates that the majority of records i labeled by the attacker as being part of the training dataset were actually part of R. In fact, any precision higher than 0.5 means that the probability of inferring that a record i within RH was part of the training dataset R is higher than

[4]Records labeled as being part of R that were actually in R.

a random 50% chance and, thus, 0.5 is used as a benchmark when defining what value of precision is deemed too high.

Privacy Hub advocates the use of the Hamming distance DCR in order to calculate the membership inference risk, as it provides arbitrary thresholds that support comparison across different percentages of R and H, and different datasets. The Hamming distance DCR thresholds can only be integers and, therefore, all viable thresholds for calculating the precision of the correct membership inference should be numbers higher than the minimum DCR but lower than the maximum DCR.

The following subsections describe membership inference testing performed on four datasets. A summary of the datasets used for testing is shown in Table 25. It is important to note that in carrying out these tests, Privacy Hub generated holdout datasets which were not shared with Subsalt, and that in using the Hamming distance DCR, numeric variables were binned. The choice of bins was based on Privacy Hub's judgment and knowledge of the usual distribution of variables in question, but might not represent the optimal choice. However, it is expected that other bins would not change the results drastically.

Table 25 – Summary of datasets used for testing.
All real datasets were split 50/50 between training and holdout data.

Scenario	Columns	Rows (real)	Rows (synthetic)	Comment on analysis
1	14	100,000	50,000	Baseline test
2	12	100	50	Designed to fail membership inference test
3	16	100,000	50,000	Multivariate relationships Effect of varying DCR threshold and % of known records
4	14	100,000	50,000	Multivariate relationships Aggregation of age

3.9.1 Testing Scenario 1

Throughout this section, the real dataset discussed and used to simulate a membership inference attack is referred to as 'Dataset 1'. The variables contained in this dataset are summarized below; the numeric variables are marked with a '*', while the remaining

variables are categorical.

*age**	*income*
state	*education*
gender	*diagnosis*
race	*temperature**
ethnicity	*oncology_stage*
death_flag	*procedure*
marital_status	*lab_result*

The dataset originally contained 100,000 rows, which were to be treated as individual patient records. However, Privacy Hub randomly split this original 'real' dataset into a holdout dataset containing 50,000 records and a training dataset containing 50,000 records. The holdout dataset was not shared with Subsalt. Once the holdout data was removed from the test dataset, the remaining 50,000 records that constituted the training dataset were sent to Subsalt. Following the delivery of this 'Training Dataset 1', Subsalt provided Privacy Hub with three synthetic datasets, one for each model (TVAE, CTGAN and CopulaGan), each of which contained the variables detailed above and 50,000 rows.

In this test, different proportions of records (100%, 50%, ..., 2.5%) from the training and holdout datasets were used to define RH, which consisted of randomly sampled and equally sized R and H subsets. The Hamming distance was used to compute the DCR between each record of the RH dataset and records within S. For the test, the *age* variable was left unchanged, matching only on exact matches between real and synthetic datasets, whilst the *temperature* variable was aggregated into 25 equal-depth bins (as defined by the real dataset) for this purpose. The results are shown in Table 26.

Table 26 – Results of membership inference test on dataset 1

		Precision		
Proportion of records	Threshold (≤)	TVAE	CTGAN	CopulaGAN
2.5%	1	0.607	0.562	0.588
2.5%	2	0.501	0.507	0.493
2.5%	3	0.501	0.496	0.507

Continued on next page

Results of membership inference test on dataset 1

		Precision		
Proportion of records	Threshold (≤)	TVAE	CTGAN	CopulaGAN
2.5%	4	0.499	0.499	0.502
2.5%	5	0.498	0.500	0.500
5%	1	0.582	0.574	0.544
5%	2	0.516	0.507	0.519
5%	3	0.500	0.496	0.506
5%	4	0.503	0.499	0.498
5%	5	0.500	0.500	0.501
5%	6	0.500	-	-
7.5%	1	0.459	0.476	0.492
7.5%	2	0.504	0.478	0.486
7.5%	3	0.495	0.498	0.493
7.5%	4	0.500	0.500	0.500
7.5%	5	0.501	0.500	0.500
7.5%	6	0.500	0.500	-
10%	1	0.525	0.470	0.493
10%	2	0.515	0.507	0.493
10%	3	0.504	0.503	0.501
10%	4	0.502	0.502	0.499
10%	5	0.501	0.500	0.500
10%	6	0.500	-	-
50%	1	0.521	0.480	0.487
50%	2	0.496	0.491	0.485
50%	3	0.501	0.499	0.498
50%	4	0.501	0.500	0.499
50%	5	0.500	0.500	0.500
50%	6	0.500	0.500	0.500
50%	7	-	-	0.500
100%	1	0.517	0.501	0.501
100%	2	0.501	0.498	0.493
100%	3	0.502	0.499	0.499
100%	4	0.501	0.500	0.499
100%	5	0.500	0.500	0.500
100%	6	0.500	0.500	0.500
100%	7	0.500	-	0.500
	Mean	0.507	0.501	0.502

On average, the precision of a membership inference attack on this dataset was very close to the random guess precision of 0.5 for each of the three models. This indicates that the synthetic datasets do not represent a membership disclosure risk to the training dataset, and that the chances of an attacker correctly inferring that a (patient) record was part of the 'real' sample are not more advantageous than those of a random guess.

3.9.2 Testing Scenario 2

The second test was conducted on the Privacy Hub-generated 'Dataset 2'. The variables contained in this dataset are summarized below; the numeric variables are marked with a '*', while the remaining variables are categorical.

*Age**	*Marstat*[5]
ZIP3	*LOINC*
Gender	*Service_date*
*Income**	*Allergies*
*Height**	*Days_in_hospital**
Race	*Facility_id*

The dataset originally contained 100 rows which were to be treated as individual patient records. Privacy Hub randomly split this original 'real' dataset into 'Holdout Dataset 2' which represented a random subset of the original dataset (containing the variables detailed above and 50 records), and 'Training Dataset 2' which contained the remaining 50 records. The holdout dataset was not shared with Subsalt. Following the removal of the holdout dataset, the remaining 50 records ('Training Dataset 2') were delivered to Subsalt, who subsequently provided Privacy Hub with three synthetic datasets, one for each model (TVAE, CTGAN and CopulaGan), each of which contained the same variables and 50 rows. This dataset was designed to fail a reasonable test of membership inference (it was expected that a synthetic dataset more similar to the training dataset than to the holdout dataset would be obtained), as the reduced size and the diversity of the records within the datasets meant that the training and holdout datasets were very different to

[5]Contained values regarding patient marital status.

start with. The 'Training Dataset 2' could be considered to be either constructed of outliers only or to contain no outliers in the real sense, as all records within the holdout and training datasets were unique and there were no particular trends between records across any of the variables.

Initially, the synthetic datasets performed somewhat similarly to the other datasets, which was contrary to Privacy Hub's expectations. On analyzing the synthetic output, it was noted that the data type of some variables did not match the real data, with some categorical variables being treated as numeric. The effect of this was that the synthetic data created a large spread of values not included in the real data, thereby creating a set of records with larger-than-expected DCRs which were then used as the basis for the membership inference calculation. This raises an important consideration both from a utility perspective and from a privacy perspective. In relation to privacy, the way in which the input data is processed in this regard may be the difference between between passing or failing a privacy test. However, the presence of such values in the synthetic data can help to mask the real values appearing in the underlying data. In such a small dataset, particularly, there is a large effect on the utility of the synthetic output. In both cases, then, there is a balance and it is important to consider the overall structure of the real dataset prior to synthetic generation. In this regard, Privacy Hub requires that variable classes for numeric and categorical variables are maintained throughout the synthetic generation process, as attributing certain variables (including demographic variables) an incorrect variable class within the synthetic data generation process can affect both the fidelity and the disclosure risk of the resulting synthetic dataset. This requirement is at the level of numeric and categorical only, and does not distinguish between data types – for example, integer and float. Of course, this test data was designed specifically with a particular privacy test in mind, with little consideration of the utility of the resulting synthetic data (it was expected that the 50-record training dataset was too small a training set to generate sensible results).

To address this, a new set of synthetic datasets was supplied by Subsalt, where two of the categorical variables – *ZIP3* and *Facility_id* – had indeed been processed as categorical by the synthetic engine. The updated membership inference results showed that the test failed as expected. The engine was overfit to the training data (as was designed, given

there were only 50 records in the training set and 100 in total). The results for this latter scenario are shown in Table 27.

The numeric variables were recoded as follows: the values of the *Age* and *Days_in_hospital* variables remained unchanged, while the values populating the *Height* and *Income* variables were aggregated into 25 equal-sized bins. Hamming distance DCR thresholds defined by the maximum and minumum DCR values for each model were used, with values lying strictly within these bounds. The results are shown in Table 27.

Table 27 – Results of membership inference test on Dataset 2

		Precision		
Proportion of records	Threshold (≤)	TVAE	CTGAN	CopulaGAN
10	6	-	-	1.000
10	7	0.556	0.571	
10	8	-	-	0.625
20	6	0.600	-	0.833
20	7	0.533	0.600	0.563
20	8	0.526	-	0.526
25	5	0.667	-	-
25	6	0.769	0.778	0.625
25	7	0.550	0.611	0.526
25	8	0.522	-	0.521
30	5	0.800	-	-
30	6	0.733	0.500	-
30	7	0.609	0.636	0.520
30	8	0.483	0.517	-
40	5	1.000	-	-
40	6	0.792	0.667	0.800
40	7	0.655	0.625	0.533
40	8	0.526	0.513	0.513
45	5	1.000	-	-
45	6	0.792	0.667	0.800
45	7	0.655	0.625	0.533
45	8	0.526	0.513	0.513
50	5	0.714	-	-
50	6	0.720	0.600	0.571
50	7	0.575	0.575	0.513

Continued on next page

Results of membership inference test on Dataset 2

		Precision		
Proportion of records	Threshold (≤)	TVAE	CTGAN	CopulaGAN
50	8	0.490	-	-
60	5	0.875	-	-
60	6	0.690	0.692	0.571
60	7	0.553	0.542	0.545
60	8	0.500	-	0.508
70	5	0.667	-	-
70	6	0.658	0.471	0.593
70	7	0.542	0.571	0.527
70	8	0.500	-	0.515
75	5	0.769	-	-
75	6	0.674	0.533	0.615
75	7	0.574	0.574	0.534
75	8	0.507	0.507	0.514
80	5	0.769	-	-
80	6	0.667	0.619	0.645
80	7	0.565	0.596	0.540
80	8	0.494	0.506	0.513
90	5	0.733	-	-
90	6	0.667	0.600	0.647
90	7	0.571	0.577	0.543
90	8	0.506	0.506	0.511
95	5	0.800	-	-
95	6	0.700	0.650	0.647
95	7	0.560	0.579	0.528
95	8	0.500	0.505	0.511
100	5	0.750	-	-
100	6	0.679	0.591	0.632
100	7	0.557	0.575	0.526
100	8	0.500	0.505	0.510
	Mean	0.641	0.579	0.583

On average, the precision of a membership inference attack on this dataset was understandably very high, as the synthetic data was overfitted to the training dataset, while the holdout dataset was not very similar to the training dataset. This is to be expected

for such a small dataset with virtually no commonality between the two randomly split training and holdout subsets. Irrespective of the threshold used or the proportion of records known to the attacker, it is apparent that the precision was generally higher than the 0.5 precision of a random guess, indicating that this dataset presents a high risk of correct membership inference.

3.9.3 Testing Scenario 3

The third test was conducted on the Privacy Hub-generated 'Dataset 3'. The variables contained in this dataset are summarized below; the numeric variables are marked with a '*', while the remaining variables are categorical. The real data contained 100,000 rows. The dataset contained multivariate relationships to better represent a real dataset likely to be seen by Subsalt. For the test, *age* was left unchanged, matching only on exact matches between real and synthetic datasets, whilst both *temperature* and *height* were aggregated into 25 equal-depth bins (as defined by the real dataset).

As before, the data was split equally into a holdout and training dataset. Subsalt generated three synthetic datasets, one for each model (TVAE, CTGAN and CopulaGan), each of which contained the same variables and 50,000 rows.

*age**	*zip3*
*height**	*plan_member*
diagnosis	*final_claim*
procedure	*death_flag*
pos	*marstat*
income	*status*
race	*ethnicity*
gender	*temperature**

The results, listed in Table 28, show that, on average, the precision of a membership inference attack on this dataset was very close to a random guess precision of 0.5.

Table 28 – Results of membership inference test on Dataset 3

		Precision		
Proportion of records	Threshold (≤)	TVAE	CTGAN	CopulaGAN
2.5	4	0.460	-	0.407
2.5	5	0.495	0.474	0.488
2.5	6	0.500	0.496	0.496
2.5	7	0.501	0.499	0.499
2.5	8	0.499	0.500	-
5	4	0.504	0.457	0.527
5	5	0.509	0.504	0.488
5	6	0.499	0.499	0.503
5	7	0.497	0.500	0.500
5	8	0.499	0.500	-
5	9	0.500	-	-
7.5	4	0.507	0.508	-
7.5	5	0.497	0.505	0.510
7.5	6	0.501	0.496	0.499
7.5	7	0.499	0.499	0.499
7.5	8	0.500	0.500	-
10	4	0.562	0.532	-
10	5	0.510	0.495	0.489
10	6	0.497	0.497	0.499
10	7	0.499	0.499	0.500
10	8	0.500	0.500	0.500
25	4	0.532	0.519	0.512
25	5	0.515	0.494	0.503
25	6	0.502	0.498	0.500
25	7	0.500	0.499	0.500
25	8	0.500	0.500	-
25	9	0.500	-	-
50	4	0.516	0.494	0.501
50	5	0.508	0.498	0.500
50	6	0.502	0.499	0.499
50	7	0.500	0.500	0.500
50	8	0.500	0.500	0.500
50	9	0.500	-	-
75	4	0.537	0.492	0.515

Continued on next page

Results of membership inference test on Dataset 3

		Precision		
Proportion of records	Threshold (≤)	TVAE	CTGAN	CopulaGAN
75	5	0.510	0.499	0.502
75	6	0.503	0.499	0.500
75	7	0.500	0.500	0.500
75	8	0.500	0.500	0.500
75	9	0.500	-	-
100	4	0.532	0.497	0.507
100	5	0.509	0.498	0.500
100	6	0.502	0.498	0.500
100	7	0.500	0.500	0.500
100	8	0.500	0.500	0.500
100	9	0.500	-	-
	Mean	0.505	0.499	0.498

Tables 29 and 30 show the mean precision broken down by threshold and by percentage of records known to the attacker, respectively. The results show that changing the proportion of real/holdout records known to the attacker does not have a big impact on the precision, as it effectively only changes the classification problem. Changing the DCR threshold is also seen to have a minimal impact on the overall performance with little variation in the mean precision. In our analyses we consider all reasonable thresholds (anything between minimum and maximum), as the purpose of our testing is to evaluate the classification's precision rather than its accuracy (which would normally be expected to yield best results for the DCR threshold closest to the average DCR).

3.9.4 Testing Scenario 4

The final test was conducted on the Privacy Hub-generated 'Dataset 4'. The variables contained in this dataset are summarized below; the numeric variables are marked with a '*', while the remaining variables are categorical. The real data contained 100,000 rows. The dataset contained multivariate relationships to better represent a real dataset likely to be seen by Subsalt, and additional consideration was given to the depth of categorical variables to ensure there was a degree of variance between training and holdout datasets.

Table 29 – Mean precision by threshold

Threshold	Mean precision		
	TVAE	CTGAN	CopulaGAN
4	0.519	0.500	0.495
5	0.506	0.496	0.498
6	0.501	0.498	0.500
7	0.500	0.499	0.500
8	0.500	0.500	0.500
9	0.500	-	-

Table 30 – Mean precision by percentage of known records

Percentage known	Mean precision		
	TVAE	CTGAN	CopulaGAN
2.5	0.491	0.492	0.473
5	0.501	0.492	0.504
7.5	0.501	0.502	0.503
10	0.514	0.505	0.497
25	0.509	0.502	0.504
50	0.504	0.498	0.500
75	0.508	0.498	0.503
100	0.507	0.499	0.501

For the test, *age* was left unchanged, matching only on exact matches between real and synthetic datasets, whilst all of *height*, *temperature*, *bmi*, *cholesterol* and *blood pressure* were aggregated into 25 equal-depth bins (as defined by the real dataset).

As before, the data was split equally into a holdout and training dataset. Subsalt generated three synthetic datasets, one for each model (TVAE, CTGAN and CopulaGan), each of which contained the same variables and 50,000 rows.

*age**
*height**
diagnosis
loinc
income
race

gender
marstat
enrolled
allergies
*temperature**
*bmi**
*cholesterol**
*blood pressure**

Once again, the results, listed in Table 31, show that, on average, the precision of a membership inference attack on this dataset was very close to a random guess precision of 0.5, demonstrating that the engine performs well under such a test even in the presence of increased variance. Of course, as described in Section 3.9.2, we have seen that in an extreme case, the membership inference test is indeed failed.

Table 31 – Results of membership inference test on Dataset 4

		Precision		
Proportion of records	Threshold (≤)	TVAE	CTGAN	CopulaGAN
2.5	4	0.462	0.385	-
2.5	5	0.515	0.479	0.538
2.5	6	0.509	0.505	0.512
2.5	7	0.504	0.501	0.499
2.5	8	0.499	-	-
2.5	9	0.499	-	-
5	4	0.469	-	-
5	5	0.506	0.493	0.523
5	6	0.496	0.500	0.501
5	7	0.498	0.500	0.498
5	8	0.498	0.500	0.500
5	9	0.500	-	-
7.5	4	0.570	0.379	0.574
7.5	5	0.530	0.491	0.511
7.5	6	0.501	0.502	0.500
7.5	7	0.499	0.500	0.500
7.5	8	0.500	-	-
7.5	9	0.500	-	-
10	4	-	0.583	-
10	5	0.521	0.506	0.495
10	6	0.501	0.495	0.500

Continued on next page

Results of membership inference test on Dataset 4

		Precision		
Proportion of records	Threshold (≤)	TVAE	CTGAN	CopulaGAN
10	7	0.501	0.499	0.500
10	8	0.499	-	-
10	9	0.500	-	-
10	10	0.500	-	-
25	4	0.491	-	0.529
25	5	0.511	0.503	0.504
25	6	0.503	0.499	0.501
25	7	0.501	0.500	0.500
25	8	0.499	0.500	-
25	9	0.500	-	-
50	4	0.509	0.477	0.538
50	5	0.510	0.504	0.508
50	6	0.503	0.499	0.502
50	7	0.501	0.500	0.500
50	8	0.500	0.500	-
50	9	0.500	-	-
75	4	0.498	0.459	0.540
75	5	0.512	0.497	0.501
75	6	0.504	0.500	0.502
75	7	0.501	0.500	0.500
75	8	0.500	0.500	-
75	9	0.500	-	-
90	3	-	0.571	-
90	4	0.500	0.489	0.533
90	5	0.508	0.492	0.501
90	6	0.503	0.500	0.501
90	7	0.501	0.500	0.500
90	8	0.500	0.500	-
90	9	0.500	-	-
100	3	-	0.571	-
100	4	0.502	0.497	0.531
100	5	0.510	0.496	0.502
100	6	0.502	0.500	0.501
100	7	0.500	0.500	0.500
100	8	0.500	0.500	-

Continued on next page

Results of membership inference test on Dataset 4

		Precision		
Proportion of records	Threshold (≤)	TVAE	CTGAN	CopulaGAN
100	9	0.500	-	-
100	10	0.500	-	-
	Mean	0.502	0.497	0.510

Furthermore, Privacy Hub conducted a second test on the synthetic data, this time considering ages aggregated into three-year bins. There was only a minor effect on the mean precision for all three of the models: both the TVAE and CopulaGAN models showed a slight decrease in the mean precision, whilst the CTGAN showed a slight increase. In each case, the difference was small and the mean precision continued to fall very close to a random guess precision of 0.5.

3.9.5 Discussion

A number of issues came to light when designing and implementing membership inference tests. Before commencing work and/or testing, consideration must be given to what should constitute an acceptably low precision level – or, more explicitly, a precision that is close enough to the 0.5 random guess precision that it does not introduce a risk of membership inference that is more than negligible. Shokri et al. (2017) report the membership attack results of six real datasets, 'attacked' using "classification models trained by commercial 'machine learning as a service' providers such as Google and Amazon". Two of the six datasets were considered to be representative of health data; Texas hospitalization data on more than 67,000 records, and census information on 50,000 individuals. The corresponding membership inference attack precision for these two datasets was reported to be 0.657 and 0.503, respectively. Thus, the mean membership inference attack precision between these two datasets was 0.580.

Having assessed the precision corresponding to a membership inference attack for four datasets, it is apparent that there is a need to define a sensible threshold below which, while there is a greater-than-random chance of accurate inference, the overall risk remains

sufficiently small and the dataset is considered safe against such an attack. Indeed, across all models, while for the testing scenarios 1, 3 and 4 the precision was close to the value of 0.5, for Dataset 2 this was not the case. The results obtained from the four tested datasets discussed in this section demonstrate that changing the proportion of real/holdout records known to the attacker does not have a big impact on the precision, as it effectively only changes the classification problem. It was also observed that changing the DCR threshold had minimal impact on the overall performance with little resulting variation in the mean precision.

Using combinations of defined percentages of records known to the attacker and Hamming distance DCR thresholds, at least 30 precision values were calculated for each of the four datasets. Using the central limit theorem (CLT), the overall precision of each of these datasets can be approximated using samples' precisions, as the precision is an estimate that is normally distributed and, as the sample of precisions calculated becomes larger, the overall precision will become approximately equal to the mean precision of the samples. Moreover, a sample of at least 30 precision values is considered sufficient for the CLT to hold. Thus, the 30 or more precision sample values calculated for each of the four datasets should be an accurate determinant of the overall precision of each of the datasets, for each model.

As a conservative approach, in order that a synthetic dataset is considered 'safe' from membership inference attacks, Privacy Hub requires that the mean precision of a dataset should be below the value of 0.550. The average precision must be calculated from at least 30 sample precision values, varying in both percentage of records known to the attacker and DCR threshold using a range of values selected between the minimum and maximum DCR and including the average threshold (to nearest integer). We consider this threshold to be reasonable while also providing a more conservative mean precision than that of the two real datasets 'attacked' and documented in Shokri et al. (2017). Furthermore, Privacy Hub recommends that the threshold should be lowered for any dataset presenting a higher level of sensitivity (e.g., a dataset containing only patients with back pain is less sensitive than a dataset containing only patients with HIV), and proposes that for such datasets the upper threshold should be 0.525.

Based on this requirement, Datasets 1, 3 and 4 are considered to be reasonably protected against a membership inference attack, while Dataset 2 presents a high risk of membership inference and is, thus, expected to fail a membership attack. Furthermore, for all three models, Datasets 1, 3 and 4 present a mean precision which is below the further recommended precision threshold of 0.525.

When constructing the *RH* dataset, Privacy Hub requires that equal proportions of the training and holdout data are used. In the examples detailed herein the real data was evenly split between training and holdout. Here we detail, by way of example, how *RH* must be constructed in scenarios where this is not the case. Let's consider that the real data is split in a 90:10 ratio of training to holdout data. When constructing *RH* using, for example, 100% of the holdout data, the training data must contribute to *RH* in the following way: firstly, the training data must be divided into nine equal sized, randomly selected and distinct subsets. In turn, each subset must be added to the holdout dataset to define *RH* for one round of membership inference testing. The results of all nine iterations of testing (one for each training subset) should be averaged. From these average results, the overall average precision can be calculated as detailed herein. Where, for example, 5% of the holdout dataset is to be used, the training data should be split into 18 randomly selected, distinct subsets and so on and so forth. In scenarios where a non-integer number of subsets of the training data is required to create RH, where it is more economical, the surplus records may simply be dropped without significantly affecting the risk, otherwise, resampling can be used to make an additional subset which contributes to RH. For example, if each subset of the training data contributes 10 records, where the final subset contains 3 surplus records, these may be dropped. Where the final subset contains 7, the remaining 3 records may be selected using resampling methods from the previous selections.

3.9.6 Requirements and Recommendations

Following analyses of the tests described in the previous sections, Privacy Hub requires and recommends the following modifications:

- Variable classes for numeric and categorical variables must be maintained through-

out the synthetic generation process.

- A dataset should be considered at small risk of membership inference if the average precision, calculated from at least 30 sample precision values, varying in both percentage of records known to the attacker and DCR threshold as detailed in Section 3.9.5, is below 0.550.
- If a dataset is deemed to be of high sensitivity, it is recommended that the threshold for membership inference precision should conservatively be reduced to 0.525.

3.10 Attribute Inference

A viable type of attack on synthetically generated data is to use the synthetically generated data to make inferences on previously unknown features of some underlying real dataset. This is known as an *inference* attack. A particular example of such an attack is *attribute inference* (see, for example, Matwin et al. (2015), El Emam et al. (2020b), Yan et al. (2020)), in which, typically, a machine learning model is applied to the synthetically generated output, in order to predict attributes of the underlying real dataset which were not previously present. Therefore, a valid privacy metric will accurately quantify the disclosure risk of the synthetic dataset, in the event of such an attack.

Recently, a novel approach to modeling the disclosure risk associated with a synthetic output under an attribute inference attack was discussed (El Emam et al., 2020a), and Privacy Hub's approach to reviewing disclosure risk for Subsalt's three model types – TVAE, CTGAN and CopulaGAN – is broadly in line with the methods described therein. Further, Privacy Hub herein provides Subsalt with recommended algorithm modifications aimed at reducing the disclosure risk in the context of attribute inference, and recommended thresholds relevant to the calculation of attribute inference.

3.10.1 Development of the Disclosure Measure

We begin by defining the set of quasi- or indirect identifiers[6] within the real dataset as Q, with $q_i \in Q$ representing the individual variables. This set includes demographic variables such as age and race along with geographic information and other identifiable information such as income, etc. We define a group of sensitive variables as V, with $v_j \in V$ representing the individual sensitive variables that an attacker seeks to learn. For simplicity, it is assumed that each record of the real dataset represents a single patient, though this simplification can easily be removed without affecting the conclusions of this report.

For a given subset of quasi-identifiers, Q_s (which represent the characteristics of the individual whose sensitive information the attacker seeks to infer), a subset of records, S_q, matching across these identifiers, is determined in the synthetic dataset. Based on this subset (S_q), the attacker aims to infer sensitive information about the target individual within the real dataset. There is an important consideration here related to the matching of patients between the real and synthetic datasets. A unique record, s, in the real data, corresponding to some specific combination of quasi-identifiers, will generally represent a larger group of matching individuals within the real population. Thus, even if a match between a real record and a synthetic record is found, the chance of identifying an individual within the real sample is affected by the number of individuals with the same quasi-identifiers that exist in the population. The method used by Privacy Hub to assign an attribute risk score currently employs a conservative approach, modeling the risk based on a successful inference of a real patient's characteristics relative to the real sample population without comparing this patient to the population at large. This latter method tends to overestimate any estimation of the risk of re-identification. In addition, there are two possible 'directions' in which we can search for a match: firstly, by choosing a real record and searching for matching records in the synthetic data, and secondly, in the opposite direction, choosing a synthetic record and searching for matching records in

[6]These are variables which, in combination, can be linked to external information to potentially re-identify individuals. They include those variables which may be readily observable, such as gender or race. On their own they may pose no risk, but in concert they may identify either individuals or small groups of individuals. As such, these variables in combination are potentially of high risk.

the real data. Although both directions are discussed in El Emam et al. (2020a), Privacy Hub has assessed disclosure risk only in the former direction, which is discussed herein.

The central idea of the method detailed in El Emam et al. (2020a) is that the disclosure risk in the event of an attribute inference attack can be quantified by calculating the percentage of patients within the real dataset for which all three conditions below are satisfied:

1. The record in the real dataset has one or more matches within the synthetic dataset based on some subset of its quasi-identifiers.
2. Based on these matches, an attacker may infer information about the sensitive variables in the real dataset.
3. These inferences are accurate and verifiable.

The disclosure risk is quantified in the following equation:

$$d_s = \frac{1}{n_s} \sum_{s=1}^{n_s} \left(\frac{1}{f_s} \cdot \lambda_s \cdot I_s \cdot R_s \right) \tag{1}$$

where:

n_s is the number of records in the real dataset

f_s is the equivalence class size in the real sample for a particular real record s

λ_s is the adjustment factor due to errors in matching

I_s is an indicator variable that takes a value of 1 if the record s in the real data has a matching record in the synthetic data, and 0 otherwise

R_s is an indicator variable that takes a value of 1 if the adversary would learn something new from matching the record s to the synthetic data outputs, and 0 otherwise.

3.10.2 Aggregation

The above equation calculates an average risk score per dataset based on a number of criteria that are explained further throughout the following sections. As detailed in the previous section, the proportion of 'risky' records is calculated, where a risky synthetic record is one which can be matched to a real record such that additional meaningful information can be gained and verified. However, while maintaining the definitions of the terms in Equation 1 and explained herein, Privacy Hub proposes the use of an alternative aggregation in determining a final threshold. We routinely consider that a patient is at a high risk of re-identification if their record in the data is in an equivalence class size less than five within the real population. Therefore, we propose that the attribute inference risk model described by Equation 1 is adjusted to represent the percentage of high-risk records within the synthetic dataset (as previously defined and detailed herein), where such a high-risk record is part of an equivalence class size of less than five within the real data. In this regard, Equation 1 is adjusted, resulting in Equation 2 described below, to include a new binary indicator variable E_s, taking a value of 1 if the synthetic record in question matches with a real record which appears in an equivalence class of less than five in the data, and 0 otherwise.

$$d_s = \frac{1}{n_s} \sum_{s=1}^{n_s} (E_s \cdot \lambda_s \cdot I_s \cdot R_s) \tag{2}$$

where:

n_s is the number of records in the real dataset

E_s is an indicator variable that takes a value of 1 if the record s matches a real record which is part of an equivalence class of less than five, and 0 otherwise

λ_s is the adjustment factor due to errors in matching

I_s is an indicator variable that takes a value of 1 if the record s in the real data has a matching record in the synthetic data, and 0 otherwise

R_s is an indicator variable that takes a value of 1 if the adversary would learn something new from matching the record s to the synthetic data, and 0 otherwise

All results related to the final proportion of risky records quoted herein were calculated using Equation 2. The following sections describe this equation in more detail and analyze its construction. Where relevant, recommendations for improvements are also included.

3.10.3 Defining a Match

The binary variable I_s captures whether or not the real record has, across its quasi-identifiers, a matching record in the synthetic data. If no match is found, the matching indicator I_s is equal to 0 and there is no risk of attribute disclosure for record s. If one or more matching records are found in the synthetic data, then I_s is assigned the value 1 for that real record. Note that at this point, any record with I_s can be effectively dropped from the calculation and cannot contribute to the disclosure risk associated with attribute inference.

For any real record with a match across quasi-identifiers in one or more synthetic records, it is important to determine the group, or equivalence class, size of the real record in question. If the real record in question appears in a group of fewer than five records with matching quasi-identifiers, then E_s takes the value 1. If the real record appears in a group of five or more, E_s takes the value 0. Thus, only records which i) have a match across quasi-identifiers with one or more synthetic records, and ii) are part of an equivalence class of less than five in the real dataset, can contribute to disclosure risk associated with attribute inference.

It is important to consider how a match is defined. In this regard, an attribute inference assessment has the potential to incorporate additional functionality to allow for 'similar' matches in near-categorical quasi-identifiers such as age, allowing for matches within a configurable margin of error, such as $\pm x$ years. This idea is discussed further in Section 3.10.9. A similar approach, where deemed relevant, can be taken for other variables, where less stringent matching criteria will produce a more conservative estimation of the risk associated with a given variable. The choice of quasi-identifiers is also important here,

and testing multiple scenarios may be advantageous. It may, in some circumstances, be pertinent to consider matches at an aggregate geographic level; for example, treating any patient within the same state as a match instead of within the same 3-digit ZIP code in order to provide a more conservative measure of risk. These recommendations should be thought of as a set of guiding principles, forming part of a general implementation of the risk model described herein, that can help to provide a more conservative measure of the disclosure risk associated with a given dataset under an attribute inference attack. There is no strict requirement to implement these measures unless explicitly stated; rather, they are recommendations and considerations which can be implemented at the discretion of Subsalt and its clients. For all results provided herein, a match is defined by an identical entry across all quasi-identifiers (i.e., no fuzzy matching).

3.10.4 Information Gain – R_s

Measuring whether an attacker will learn something new if record s in the real sample matches a record in the synthetic sample is incorporated in the binary indicator R_s. This variable is dependent on I_s as to whether or not something is learned from a match, and is innately dependent on whether or not there is a match across the quasi-identifiers.

In the case in which the attacker is able to learn something new for a matched record, the degree to which this is sensitive information influences the choice between the values of 0 and 1 in the R_s indicator. The method incorporates whether an adversary learns something new by two criteria:

1. How similar is the real patient's information, relative to the rest of the real population? I.e., to what extent is the individual discernible within the real sample?
2. How similar is the synthetic sample value to the real sample value?

If, for example, the information gained relates to the height of an individual, and that individual sits comfortably around the average height value within the data, or within some densely populated subgroup, the information gain here is unlikely to dramatically increase the risk of disclosure. However, if the individual is a unique or close to unique

outlier within the data, with an extreme value for height, then the information gained may make the individual highly identifiable. In this second scenario, in particular, the proximity of the synthetic value to the real value becomes important. If the synthetically generated height is even within a few centimeters of an extreme value, it may still serve (particularly in combination with other quasi-identifiers) to make the individual readily identifiable, or at least to narrow the range of possibilities to a small group.

The question of how to determine the similarity of an attribute is, of course, a nuanced one and it involves the idea of distance between records or variables. In El Emam et al. (2020a), two methods are outlined for calculating this distance for nominal and binary sensitive variables and continuous sensitive variables, respectively. We briefly detail these below.

Nominal and Binary Sensitive Variables

For the binary sensitive variable X_s, we define J as the set of values that X_s can take in the real sample. Assuming that $X_s = j$ with $j \in J$ we define p_j as the proportion of real records that take the value j in this variable.

The distance between values is then defined as

$$d_j = 1 - p_j \tag{3}$$

This value is multiplied by an Iverson bracket to account for whether there is a match between the real and synthetic values Y_t and X_s, respectively.

To place a threshold on whether or not the adversary has learned something new, a conservative bound of one standard deviation (assuming a Bernoulli distribution of values) is used:

$$d_j x \left[X_s = Y_t\right] > \sqrt{p_j(1 - p_j)} \tag{4}$$

The information gain for categorical variables is determined in a similar fashion, as categorical variables are transformed into dummy variables in order to reduce the calculation

to a series of binary decisions, assuming a Bernoulli distribution on each. Thus, for a sensitive variable split into n categories, the distance is dependent on the number of real records with a specific value, Y_t.

Inequality 4 captures both the similarity between the real value and the synthetic value as well as the similarity between each real patient's information and that of the rest of the patients/records. For instance, in the case of class imbalance of a binary sensitive variable, assuming there is a match between the real and the synthetic value, only a match on the rare value will be considered risky. Privacy Hub believes that while this is a sensible approach in determining how discernible a sensitive variable's value is from all the other potential values present within the real dataset, improvements can be made in order to capture the difference in sensitivity between levels of the same nominal or binary sensitive variable. One possible such improvement is the following augmentation in order to account for different levels of sensitivity: where one (or more) of the values populating a sensitive variable is largely understood to be carrying more sensitive information than the other values, the threshold that the distance $d_j x$ is compared to is lowered (e.g., to 0.5 $\times$ standard deviation). Additionally, where there is a match between the real and synthetic values of the sensitive variables, and this value is not in fact disclosing specific sensitive information, the attribute disclosure of the record need not be considered in the attribute inference risk of the overall dataset. Here we refer to values such as 'Other' or 'Unknown' or missing values.

Continuous Variables

For continuous variables, El Emam et al. (2020a) recommends discretizing the data using univariate k-means clustering and choosing optimal cluster sizes using a majority rule. Privacy Hub, on the other hand, discretizes the data in the same way but chooses optimal clusters based on the Elbow method.

In a similar fashion to that described above for discrete variables, the distance is defined as the proportion of patients occurring in the given cluster relative to all patients in the real dataset.

$$d_S = p_s \tag{5}$$

Defining the weighted absolute difference as $d_S \times |X_s - Y_t|$, we can set the threshold as

$$d_s \times |X_s - Y_t| < 1.48 \times MAD \tag{6}$$

where MAD is the median absolute deviation and is normalized to one standard deviation for Gaussian distributions (El Emam et al., 2020a). This inequality captures the trade-off between the relative frequency of the cluster to which the value belongs and the distance between the real value and the synthetic value. In this way, records for which the real and synthetic values are very similar will not be considered risky if there is a high enough proportion of patients whose real values within this variable are part of the same cluster.

Privacy Hub advocates accounting for both the difference between the real and the synthetic values, as well as for the cardinality of the real record's cluster. In order to be consistent with the way in which categorical variables are treated, Privacy Hub utilizes a cluster-specific approach to calculating the MAD, in order to compare the weighted absolute difference $d_s \times |X_s - Y_t|$ to the MAD, calculated on a cluster-by-cluster basis. Take, for example, a variable containing the white blood cell count of a patient (note that direct inference of values in this variable will not constitute a privacy breach; however, it can represent valuable information gain in an attribute inference attack): in the case that two clusters are equally small (for example a cluster of four real values representing number of white blood cells: 2, 5, 16 and 17, and a cluster of four real values representing number of white blood cells: 50, 52, 55 and 60) and have very different distributions, all else being equal, the values within the cluster with a larger MAD (6.5 vs. 3.25 in our example) will have more chances to satisfy the inequality.

High-Risk Records

If the attacker is able to accurately infer new information on some attribute for a real record, s, and it is determined by the previously described thresholds that the information gain is valuable, then the indicator variable, R_s, is set to 1 and the variable is deemed high risk. If not, the variable is deemed safe and R_s is set to 0. Thus, the question of how many sensitive values a row can contain before it is deemed to contain too much risk arises. In the El Emam et al. (2020a) paper, the authors set a threshold of $L\%$ of sensitive features being deemed high risk before a row is deemed high risk. However, Privacy Hub has implemented a more conservative approach, effectively setting $L = 0$,

translating into: if any sensitive variable is high risk, then the record is deemed to be high risk. This approach is designed to capture more potential sensitive information disclosure that may be overlooked in the case in which the 'high-risk' sensitive variables collectively fall below the $L\%$ threshold but which do still significantly increase the disclosure risk.

3.10.5 Sensitive Variables

There is an inherent ambiguity in what constitutes a 'sensitive variable'. Moreover, within any reasonable estimation, there is additional ambiguity in how different sensitive variables should be treated and a scale of the associated contribution to disclosure risk. Clearly, in a practical setting, the risk of patient re-identification is drastically higher for certain variables than it is for others. The method described in this report depends implicitly on both the choice of quasi-identifiers and the sensitive variables which may be classed as 'at risk'. A more thorough approach to managing the disclosure risk relies on addressing this dependence.

There are a number of options for addressing this issue, each of which may be considered as part of a more general implementation of the disclosure risk model described herein. It is possible to allow for individual configuration of the variables upon ingest into the synthetic engine. With a weighting parameter, defaulting to 1, assigned to each sensitive variable, it is possible to manually increase the risk associated with a given variable, if deemed necessary. This allows Subsalt and its clients to configure stricter risk thresholds for highly sensitive information or where deemed necessary due to commercial considerations, where a more or less risk-averse strategy is required. This additional weighting may also be absorbed within other, existing thresholds. For example, one may consider lowering the acceptable risk threshold, τ, for particularly sensitive variables. We may also account for sensitive variables at the aggregate level, demanding that no high-risk records containing particular sensitive variables are included in the final dataset. These ideas may also be extended to combinations of sensitive variables. The current method employed by Privacy Hub considers all sensitive variables equally and, therefore, uses the threshold values defined in Equations 4 and 6.

3.10.6 Disclosure Risk Threshold

In seeking to place measurable thresholds on the disclosure risk associated with an attribute inference attack, there are a number of threshold calibrations to consider:

1. Firstly, there is an inherent threshold within Equation 2 related to the risk associated with a sensitive variable, which translates to under what circumstances a sensitive variable is considered high risk. The approach of comparing the frequency of the specific sensitive value (or cluster for continuous variables) to the (approximated) standard deviation of the variable class (or cluster) is considered sensible, subject to further augmentation of this threshold for either very sensitive values of binary/categorical variables or extreme clusters/values of continuous variables.

2. Secondly, one can aggregate to the level of data row. How many sensitive variables may a row contain before the row itself is considered high risk? Throughout this report, this has been captured in the $L\%$ threshold. The conservative approach here is a straightforward one: set a threshold of $L = 0$ such that high-risk rows are those which contain one or more high-risk variables. Privacy Hub currently implements this $L = 0$ threshold.

3. Thirdly, there is aggregation at the level of the dataset. How many rows may be classed as high risk, as defined by use of the aforementioned thresholds, before the dataset as a whole is deemed to contain an unacceptable level of risk?

Determining a suitable disclosure risk threshold for point 3 above is not straightforward. While a general consensus has evolved under the HIPAA guidelines for real data, no such consensus has been reached for synthetic data. More generally, across many regulatory boards, a common measure of disclosure risk is centered around the risk associated with the equivalence class size of aggregated data, with records being classed as high risk should they form part of an equivalence class with cardinality less than some pre-defined threshold. The generally accepted measure of risk, which has developed around the HIPAA guidelines for real data, is that less than 1% of records may be deemed high risk, where a high-risk record is one that occurs in an equivalence class size of less than five.

Although both real and synthetic datasets carry risk, the most effective method of quantifying this risk differs between the two types. Generally speaking, risk within a real dataset can be quantified in relationship to the real population, whereas making such comparisons is more difficult when considering a synthetic dataset. While re-identification risk of a real dataset may be assessed in relation to the underlying population, the appraisal of risk within a synthetic dataset assumes the population is limited to the dataset. That is, risk is assessed relative to the only values in the data. This difference ought to be taken into account with a less stringent threshold being placed as an overall measure of disclosure risk. However, a cautious approach to implementing such thresholds is still advisable while a general industry-wide consensus develops.

There is a wide degree of variation amongst regulatory bodies on what constitutes an acceptable threshold. For example, NHS Scotland has defined different thresholds for the minimum equivalence class size for sensitive and non-sensitive data, setting it at three and either five or ten, respectively. Further examples of regulatory thresholds may be found in El Emam et al. (2020a) and elsewhere in the literature. The European Medicines Agency (EMA) has established a policy of implementing a maximum risk threshold of 0.09 (European Medicines Agency, 2019), seemingly corresponding to an equivalence class size of 11, on the publication of clinical trial data. Health Canada has implemented the same threshold for the sharing of clinical trial data.

As a parallel to the re-identification risk assessment of real data via a Privacy Hub Expert Determination, we maintain the use of an equivalence class size of less than five as a threshold. This threshold seems appropriate in the context of the previously defined attribute inference framework. As described above, there is regulatory standing for the use of the threshold, and it represents a conservative choice on the scale of regulatory advice. However, Privacy Hub considers two important and distinguishing factors when considering the overall threshold of allowed risk. Firstly, a successful attribute inference attack does not definitely imply a disclosure breach in relation to patient re-identification; rather, it is the information gain of a sensitive variable(s) that is achieved. Secondly, the equivalence class size (f_s) used in the calculation of disclosure risk in Equation 1 refers to that of the real dataset rather than that of the real population (used in Privacy Hub's Expert Determination approach) which is likely to be larger and, therefore, the risk may

be overstated when calculated in this way. It is evident, then, that the threshold for the overall percentage of high-risk records within a dataset, in relation to an attribute inference attack, can be higher than the analogous 1% used by Privacy Hub in a real data Expert Determination. Furthermore, the datasets analyzed herein are considered to contain a relatively small number of records (based on Privacy Hub's knowledge of the expected size of a health dataset), and as such, the equivalent classes within these datasets are expected to be smaller than those of most health datasets. As such, Privacy Hub requires that at most 10% of synthetic records are deemed high risk as defined herein and described by Equation 2. However, we recommend the use of a lower threshold of 5% to ensure that for reasonably sized datasets, the risk of a successful attribute inference attack is further mitigated, particularly while technology and industry consensus develops. Privacy Hub periodically reviews this 10% required threshold as knowledge of attribute inference attacks evolves and as further evidence by way of testing is gathered.

3.10.7 Error Estimates

In all analyses discussed so far, it has been implicitly assumed that all information gained from attribute inference is 100% accurate and verifiable. This is equivalent to setting $\lambda_s = 1 \ \forall \ s$ in Equation 2. Following the approach of El Emam et al. (2020a), one can account for more realistic estimations of errors in data. This information is captured in the λ_s term and accounts for errors introduced by inaccuracies in health data and for the probability that it is not feasible to verify whether a suspected match (between a real record and a record in the synthetic sample) can be verified. The two error estimates captured by λ_s are:

- A verification success rate, $\alpha = 23\%$, reflecting the proportion of times an attacker can successfully verify gained information.
- A calculated weighted mean error rate, $\beta = 4.26\%$, for health-related data.

In El Emam et al. (2020a), these terms are used in defining $\lambda_s = \lambda = 0.23 \times (1 - 0.0426)^k$, where k is the number of quasi-identifiers. Although this equation has empirical justification, Privacy Hub recommends using values of $\alpha = 1$ and $\beta = 0$, representing data

which is 100% accurate and verifiable. Additionally, the accuracy of the values derived in El Emam et al. (2020a) can be affected by technological improvements, estimation bias and available resources for such an attack.

3.10.8 Attribute Inference Results

This section presents the attribute inference results, calculated using Equation 2, for four distinct datasets. For each dataset, we present the result of Equation 2 for each algorithm type used by Subsalt: TVAE, CTGAN, and CopulaGAN. Results are provided as a percentage of records in the dataset, by multiplying the result of Equation 2 by 100. Additionally, we provide the percentage of records with $I_s = 1$, $E_s = 1$ and $R_s = 1$ for each algorithm type to facilitate discussion of anomalies specific to each algorithm.

Dataset 1

Dataset 1 comprised 100,000 records containing the variables listed below. Those variables defined as quasi-identifiers are marked with an asterisk (*), and those defined as sensitive within the calculation are marked with a dagger (†). This dataset was the same as Dataset 3 used in the membership inference testing (Section 3.9), which contained multivariate dependencies built to simulate a dataset consistent with that seen in reality.

*age**
height†
diagnosis†
procedure†
pos†
income†
*race**
*gender**
*zip3**
plan_member
final_claim
*death_flag**
*marstat**
*status**
*ethnicity**
temperature

Three synthetic datasets, each consisting of 50,000 records, were generated. Note that these were produced using the training dataset (see Section 3.9); thus, they contained fewer records than the full dataset. The algorithm type and associated attribute inference

score for each of these datasets is shown in Table 32.

Table 32 – Dataset 1: attribute inference scores for different algorithm types. Numerical values indicate the percentage of real records that equal 1 for the given metric.

Type of algorithm	I_s	E_s	R_s	d_s
CTGAN	10.8	96.7	9.2	8.1
TVAE	6.6	96.7	6	5.1
CopulaGAN	9	96.7	6.5	5.8

As is clear from the table, each algorithm produced results that fell below the 10% threshold and, thus, passed the attribute inference test. However, the results were above our recommended threshold of 5%. The high percentage of $E_s = 1$ is consistent with a large number of quasi-identifiers. Generally, more identifiers mean each individual is specified to a finer precision, resulting in a decrease in the average equivalence class size, and a corresponding decrease in the number of records that are part of an equivalence class size of five or greater.

Dataset 2

Dataset 2 comprised 100,000 records containing the variables listed below. In this case, the number of quasi-identifiers was reduced to four, and all indications of patient geography were removed. Note that this dataset was the same as that used in Section 3.9. Three datasets, each of 50,000 records due to splitting of the data into training/holdout, were generated.

*age**
height†
diagnosis†
loinc†
income†
*race**
*gender**
*marstat**
enrolled
allergies
temperature
bmi†
cholesterol†
blood pressure†

Table 33 – Dataset 2: attribute inference scores for different algorithm types

Type of algorithm	I_s	E_s	R_s	d_s
CTGAN	92.2	10.5	92.2	5.3
TVAE	58.2	10.5	58.2	0.05
CopulaGAN	91.4	10.5	91.4	5.5

Again, each algorithm fell below the required 10% threshold, with the TVAE algorithm also meeting our 5% recommendation. Contrary to Dataset 1, the small number of quasi-identifiers resulted in a low number of records with $E_s = 1$. Note that the TVAE attribute inference score is two orders of magnitude smaller than both the CTGAN and CopulanGAN scores. This is due to a smaller overlap between records where $I_s = 1$ and $E_s = 1$. This indicates that the algorithm was generating a high number of records with quasi-identifier combinations which appeared in an equivalence class of a size greater than or equal to five in the real data. Conversely, the algorithm was generating a very low number of records (0.05%) with quasi-identifier combinations consistent with real records in an equivalence class size of less than five.

Dataset 3

Dataset 3 consisted of 10,000 records, which were used to generate three further sets of the same size. This set was similar to Dataset 2 in number of assigned quasi-identifiers. Note that in this case, state was considered a sensitive variable. An important feature of this dataset is that there was missing data within the sensitive variables. To deal with this, we assumed any null entries to be distinct and meaningful. Thus, records with two matching quasi-identifiers and a null entry in the third were assumed a match. This represents the most conservative approach.

*age**
state†
*gender**
*race**
height_in†
diagnosiscode†

Despite having only three defined sensitive variables, this dataset resulted in a high num-

Table 34 – Dataset 3: attribute inference scores for different algorithm types

Type of algorithm	I_s	E_s	R_s	d_s
CTGAN	96.8	9.7	94.8	5.3
TVAE	90.4	9.7	88.1	1.2
CopulaGAN	96.2	9.7	88.4	4.7

ber of records with $R_s = 1$. This is due to state being interpreted as a sensitive variable. This introduced a categorical sensitive variable with 50 possible entries. Thus, for a given real record with quasi-identifier matches in the synthetic data, any change in the state variable (to one of 49 other states) results in $R_s = 1$ for that given real record. This results in a strong correlation between I_s and R_s (which is always a subset of I_s due to the way in which it is defined and calculated), as can be seen. However, despite a high R_s percentage, all algorithms fell below the 10% threshold, with the TVAE and CopulaGAN algorithms falling below our recommended 5% threshold.

Dataset 4

Dataset 4 consisted of 2,085 records, from which three synthetic sets of the same number of records were generated. Within this set, seven variables were considered quasi-identifiers. Due to the number of possible quasi-identifier combinations, all records in this dataset were part of an equivalence class of a size less than five; thus, $E_s = 1 \ \forall \ s$. The relationship between state and ZIP was accounted for in the generation of the synthetic datasets. Ignoring such relationships can reduce data utility (meaningless ZIP–state combinations may miss important characteristics of the data) and can skew the attribute inference score by altering the match rate between real and synthetic records. Null entries were present across quasi-identifier and sensitive variables – these were all considered distinct entries, consistent with the approach taken for Dataset 3.

*age**
*race**
*gender**
*state**
*death_date**
*marital_status**
*Zip**
HIV†
blood_group†
diagnosiscode†
white_blood_cells†

Table 35 – Dataset 4: attribute inference scores for different algorithm types

Type of algorithm	I_s	E_s	R_s	d_s
CTGAN	2.4	100	1.6	1.6
TVAE	5.7	100	3.9	3.9
CopulaGAN	1.7	100	1.0	1.0

As shown in Table 35, all algorithms produced datasets which fell below the attribute inference threshold. This was mainly due to a very small percentage of real records having $I_s = 1$. From the seven quasi-identifiers listed above, there was a total of 28,299,456 possible combinations of quasi-identifiers. As the data provided had only 2,085 records, there was a high likelihood that one of the seven variables would be altered in the synthetic dataset, resulting in a low likelihood that a match would be found. However, the number of records with a match in the TVAE dataset was almost double that of the other two datasets. On further investigation, it was found that the TVAE algorithm was poorly representing race categories. In the real dataset, *race* was grouped into seven categories, shown in Table 36. While these seven categories were fairly represented in the CTGAN and CopulaGAN schemes, only 'White' (99.8%) and 'Black' (0.2%) were present in the TVAE dataset. Thus, approximately all those real records with 'White' race would have a match in the TVAE dataset on that quasi-identifier. This resulted in a notably high TVAE attribute inference score.

Table 36 – Percentage of race values in Dataset 4

Race	Percentage of records
White	59.6
Hispanic	18
Black	12.9
Asian	5.4
Multiple Races	< 1
American Indian/Alaska Native	< 1
Samoan	< 1

To investigate the influence of the *race* identifier, the attribute inference score was calcu-

lated for the TVAE dataset ignoring patient race. It was found that this increased the risk score to 6.2%. In removing the *race* variable from the calculation, one quasi-identifier was removed, resulting in a higher match-rate (9.7%) between real and synthetic records, and leading to a higher overall risk score of 6.2%.

Upon receiving these results, Subsalt reconfigured the TVAE algorithm used to generate the synthetic data and provided Privacy Hub with a new dataset which, as shown in Table 37, more accurately represented the distribution of race within the real data.

Table 37 – Distribution of race values in the second TVAE dataset, based on Dataset 4

Race	Percentage of records
White	71.2
Hispanic	13.8
Black	10.7
Asian	1.8
Multiple Races	1.6
American Indian/Alaska Native	< 1
Samoan	< 1

Using this dataset resulted in an attribute inference score of 2.7, which is less than that previously calculated. This is consistent with the introduction of a more granular set of races in the synthetic dataset, leading to a reduction in the proportion of real records with $I_s = 1$.

Finally, to investigate the effect of ignoring ZIP–state restrictions, Subsalt provided Privacy Hub with three synthetic datasets that did **not** account for these restrictions in the generation process. The results showed a change of -0.5%, -0.4% and $+0.1\%$ for the CTGAN, TVAE and CopulaGAN algorithm attribute inference scores, respectively. This demonstrates the importance of taking such limitations into account: for the first two algorithm types, the match rate between real and synthetic records was reduced due to the presence of impossible ZIP–state combinations in the synthetic data. Note that the CopulaGAN algorithm exhibits an increase in attribute inference score. This demonstrates that, despite the general tendencies outlined above, each model accounts for the

dependencies within real data in a unique way.

3.10.9 Defining a Match II – Similarity to Real Record Testing

We return now to the discussion in Section 3.10.3 and the definition of a matching record. In particular, in order to further explore the implications of what can be considered successful information gain, it must be considered whether the information gain should only be calculated for an exact match between the quasi-identifiers of a real record s and corresponding record(s) in the synthetic dataset. To this end, Privacy Hub has assessed the case where slight variations in numerical and date–time quasi-identifiers within the dataset are also considered a match. Doing so can help diminish unreported attribute inference risk where the engine-generated records are near identical to real quasi-identifier combinations, which are overlooked in an exact match setting. In order to test the effect of this 'fuzzy' matching, Privacy Hub has calculated the attribute inference risk for Dataset 4 in each of the following scenarios:

1. Synthetic matches are determined by identical values in all quasi-identifiers.
2. Synthetic data matches are determined by a +/- 1-year shift in age and a +/- 14-day shift in death date, with identical matches in all other quasi-identifiers.
3. Synthetic data matches are determined by a +/- 2-year shift in age and a +/- 14-day shift in death date, with identical matches in all other quasi-identifiers.
4. Synthetic data matches are determined by a +/- 3-year shift in age and a +/- 14-day shift in death date, with identical matches in all other quasi-identifiers.

Note that for the TVAE algorithm, the dataset with corrected race representation was used. Results for the attribute inference score under each matching scenario are shown in Table 38.

As indicated by these results, loosening the criteria on which a match is based results in a higher attribute inference score due to an increase in the number of real records assigned $I_s = 1$. The increase was particularly significant in the TVAE algorithm, demonstrating

Table 38 – Dataset 4: attribute inference scores for different match conditions, presented for each algorithm type

Match condition	CTGAN	TVAE	CopulaGAN
Exact	1.6	2.7	1.0
Age ±1; death_date ±14 days	4.4	9.5	3.7
Age ±2; death_date ±14 days	7.6	14.3	6.0
Age ±3; death_date ±14 days	9.7	18.2	8.2

that age variables are only slightly altered in the synthetic generation process. Of course, in reality the choice of match criteria must strike a balance between being meaningful and reasonably conservative. For instance, increasing the age match criteria to ±5 could be considered overly conservative. Upon further investigation of the fuzzy matching analysis, it was confirmed that loosening the age-match criteria was the main contributor to increasing the attribute inference score; fuzzy matching on death date made little to no difference. This is due to the number of 'null' entries in the *death_date* variable, which, as discussed above, are considered distinct categorical entries, resulting in a high number of exact matches.

3.10.10 Requirements and Recommendations

Following analyses of the tests described in the previous sections, Privacy Hub requires and recommends the following modifications:

- Privacy Hub requires that at most 10% of synthetic records are deemed high risk as defined herein and described by Equation 2.
- Privacy Hub recommends the use of a lower threshold of 5% to ensure that for reasonably sized datasets, the risk of a successful attribute inference attack is further mitigated.
- Privacy Hub recommends using values of $\alpha = 1$ and $\beta = 0$ in Equation 2, representing data which is 100% accurate and verifiable.

3.11 Summary of Requirements and Recommendations

Within this section, the requirements and recommendations made throughout the report are summarized within six distinct categories: i) general actions required (their implementation applies to the synthetic engine at large); ii) dataset-specific actions required (determined on a dataset-by-dataset basis); iii) required thresholds; iv) general actions recommended; v) dataset-specific actions recommended; and vi) recommended thresholds.

General Actions Required:

1. Direct identifiers should continue to be redacted from the real data before the synthetic data generation, as detailed in Section 3.2.1.
2. Subsalt should continue to ensure that if a column within the real data is found to contain only one value, that column is discarded from the synthetic output, as detailed in Section 3.2.12.
3. Subsalt should continue to ensure that any free-text variables within the real data are either redacted or fully assessed to determine that no PII is contained within them, prior to the synthetic data generation, as detailed in Section 3.2.15.
4. Privacy Hub requires that very small datasets are not used for synthetic generation as detailed in Section 3.3.
5. As regards a recurring feed of data, as detailed in Section 3.6, where the entire dataset is to be regenerated at each update, Subsalt must require that all older versions of the synthetic data are removed to ensure that statistical inferences cannot be made against a growing number of synthetic datasets generated from similar real datasets. Privacy Hub assumes that Subsalt has appropriate data use agreements with its clients in this regard.
6. As regards a recurring feed of data, as detailed in Section 3.6, where only the additional records, making up the update, are synthesized, Privacy Hub requires that both the synthetic records generated from the update, and the fully updated synthetic output (consisting of the original synthetic output with all appended synthetic updates), are assessed using all tests as detailed in Sections 3.7 and 3.10.

7. Variable classes for numeric and categorical variables must be maintained throughout the synthetic generation process.

Dataset-Specific Actions Required:

1. Duplicate records should be removed from the real data prior to the synthetic data generation, as detailed in Section 3.2.3.

2. It is required that where dates of birth/death can be inferred from combinations of dates and codes/values implying birth/death, these are considered when evaluating the risk in relation to DCR, membership inference and attribute inference, as detailed in Sections 3.2.5, 3.7, 3.9 and 3.10.

3. For records with a DCR = 0, either the associated residential facility or code/value implying residency should be removed from the synthetic record, as detailed in Section 3.7.

4. Where a synthetic record(s) exists that contains exactly the same indirect identifiers as a real record(s), and if the corresponding real record is part of an equivalence class of five or fewer, either the associated residential facility or code/value implying residency should be removed from the synthetic record(s), as detailed in Section 3.2.10.

5. Privacy Hub requires, as detailed in Section 3.2.13, that a new threshold is defined in conjunction with the 'subsalt_other' mechanism, such that, in scenarios where only one value falls below κ_i in a given column, the 'subsalt_other' obfuscation is not used.

6. A minimum threshold of 5% of records are used as a holdout dataset which can be used for testing as detailed herein.

7. The distribution of the DCR between the real (training) dataset and the synthetic dataset should not be significantly different to that between a holdout dataset (randomly selected subset of the real dataset, not used for training) and the synthetic dataset, as detailed in Section 3.7.

8. As described in Section 3.8, duplicate records in the synthetic dataset, which are derived from unique records in the real dataset, should be removed from the synthetic dataset. A single occurrence of the record may be maintained in the synthetic data only if the record exists within a group size, as defined by the indirect identifiers, larger than five within the real data.

9. The *RH* dataset used for membership inference testing must contain equal proportions of both the training and holdout datasets as detailed in Section 3.9.

Required Thresholds:

1. As described in Section 3.7, the number of records within the real dataset that have a DCR = 0 in relation to records in the synthetic dataset, and that are in an equivalence class of five or fewer, should be lower than 1% of the total number of records within real dataset.

2. As described in Section 3.8, for large equivalence classes, which are outlier groups with regard to their group size in the synthetic data, the unique patients forming the group must exist within an equivalence class larger than five within the real data. Otherwise, the records must be removed from the synthetic data.

3. A dataset should be considered at small risk of membership inference if the average precision of a membership inference attack, as detailed in Section 3.9, is below 0.55.

4. Privacy Hub requires that, at most, 10% of synthetic records are deemed high risk as defined by Equation 2 within Section 3.10.

General Actions Recommended:

1. It is recommended that Subsalt continues to redact unique patient identifiers before the real dataset is processed by the synthetic engine.

2. A future review is recommended if the way that dates are processed (as detailed in Section 3.2.5) changes.

3. Privacy Hub recommends, and indeed assumes, that in order to mitigate the risk when sharing synthetic datasets created from the same real dataset with *different*

clients, a DUA will be in place between Subsalt and any individual recipient of the synthetic output. See Section 3.1 for more details.

4. Subsalt should continue to ensure that the number of datasets generated from a single dataset or a group of similar datasets remains low enough, such that statistical inferences cannot be made against the synthetic engine and the associated synthetic output. See Section 3.1 for more details.

5. For date of birth, it is recommended that a real and a synthetic record are considered to match if the two values match on the calendar year.

6. For date of death, it is recommended that a real and synthetic record are considered to match if the two values match on year/month where only year/month of death is present in the dataset, or if they are within the same 14 day period where date of death is present in the dataset.

7. Privacy Hub recommends, that a real training dataset (which may represent either all or a subset of the input data) contains a minimum of 3,000 records as detailed in Section 3.3.

Dataset-Specific Actions Recommended:

1. Where the size of the real dataset allows, and where utility will not be unduly affected, as much as 50% of records should be used to form a holdout dataset. The proportion of records included in a holdout dataset may vary depending on the size of the dataset. Privacy Hub recommends that a minimum of 10% of records are used, except in scenarios where the real data is very small (of the order of a few thousand records). We allow discretion in this choice, setting a required lower threshold of 5% of records.

2. Privacy Hub recommends that where the size of the real dataset is small or very small, the holdout dataset used to evaluate the similarity between the real and synthetic datasets (see Section 3.7) should represent less than 50% of the total records. Training dataset, R, should contain the same number of records as the holdout dataset, H, and should be randomly selected from the remaining records of the real dataset following the extraction of the holdout data. Testing of the

two resulting DCR distributions (DCR(R, S) and DCR(H, S)) should be repeated multiple times, with different records included within training dataset R, to ensure that the results are consistent and free of sampling bias.

3. When calculating the DCR using the Hamming distance, numerical variables should be binned based on their relevance and sensitivity rather than applying a blanket '25-bins' approach. Privacy Hub recommends that for indirect identifier variables such as age or year of birth, exact year values are used, while for other variables such as height, weight, income etc., between ten and 25 bins are used, depending on the spread of the variable.

4. Privacy Hub recommends that when considering the risk of an attribute inference attack, age related variables are matched on a range of age ±2 or 3 years, as detailed in Section 3.10.

Recommended Thresholds:

1. As described in Section 3.8, for large equivalence classes, which are outlier groups with regard to their group size in the synthetic data, the unique patients forming the group must exist within an equivalence class larger than five within the real data. Otherwise, the records must be removed from the synthetic data.

2. Privacy Hub recommends that if a dataset is deemed to be of high sensitivity, the threshold for membership inference precision should conservatively be reduced to 0.525.

3. Privacy Hub recommends the use of a lower threshold of 5% in order to ensure that for reasonably sized datasets, the risk of a successful attribute inference attack (see Section 3.10) is further mitigated, particularly while technology and industry consensus are developing.

4. Privacy Hub advises that the 10% threshold mentioned in Section 3.10 is further reviewed as the knowledge of attribute inference attacks evolves and as further evidence, by way of testing, is gathered.

4 Summary

Subsalt's process for the generation of synthetic data has been analyzed in sufficient detail to consider its alignment to the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The process has been summarized within this report and outcomes of the analyses presented. Conclusions have been drawn and documented throughout Section 3, with consideration of the HIPAA Privacy Rule and the disclosure risk of datasets included within the reported process.

4.1 Statement on Findings of Assessment

Privacy Hub finds that the synthetic data generated by Subsalt does not impart significant risk of disclosure to individuals whose real data was used as the basis for the synthetic data. In relation to HIPAA, the disclosure risk for a synthetic dataset generated by the Subsalt process is, therefore, sufficiently low to be termed 'very small', provided the requirements stipulated in Section 3.11 are implemented. Privacy Hub is confident that the proposed metrics and thresholds accurately measure disclosure risk, and that the requirements detailed herein are effective in reducing the disclosure risk of the synthetic data produced by Subsalt's engine.

4.2 Stipulated Conditions

This section summarizes the important points in reaching a conclusion regarding the status of the Subsalt process described herein, with respect to HIPAA recommendations. These conditions constitute both constraints that must be met to be consistent with HIPAA regulations, as well a true reflection of the status of the process as it stands.

This report's validity is conditional upon continued adherence to the following points:

- The process described within this report is followed as documented.
- The assumptions documented herein, upon which this report is conditional, remain true.

- Subsalt has adequate security arrangements in place, consistent with the HIPAA Security Rule.
- All requirements, as detailed throughout Section 3 and summarized in Section 3.11, are implemented. Please note that we refer here strictly to requirements (excluding recommendations); although recommendations are also listed in Section 3.11, these are not mandatory.
- The methods and technology used herein remain current and valid and are not deemed to be outdated or superseded.

References

El Emam, K., Mosquera, L., and Bass, J. (2020a). Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. *Journal of Medical Internet Research*, 22(11).

El Emam, K., Mosquera, L., and Hoptroff, R. (2020b). *Practical Synthetic Data Generation*. O'Reilly.

European Medicines Agency (2019). European medicines agency policy on the publication of clinical data for medicinal products for human use: Ema/144064/2019.

Federal Committee on Statistical Methodology (2005). Statistical policy working paper 22: report on statistical disclosure limitation methodology, 2nd version.

Matwin, S., Nin, J., Sehatkat, M., and Szapiro, T. (2015). A Review of Attribute Disclosure Control. In: Navarro-Arribas G., Torra V. (eds) Advanced Research in Data Privacy. Studies in Computational Intelligence, vol 567.

Shannon, C. E. (1949). Communication theory of secrecy systems. *The Bell System Technical Journal*, 28(4):656–715.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy*.

Yan, C., Zhang, Z., Nyemba, S., and Malin, B. A. (2020). Generating electronic health records with multiple data types and constraints. *AMIA Annu Symp Proc*.

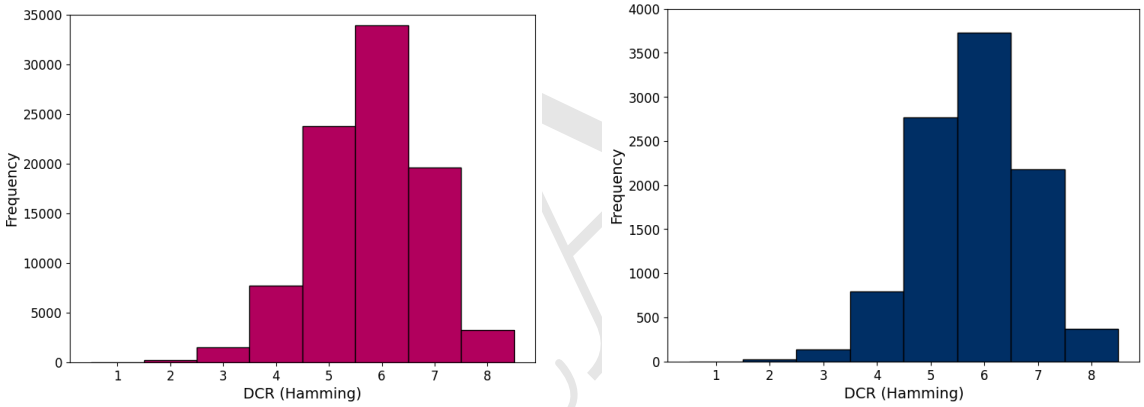
Appendix A DCR Plots

This appendix shows all relevant DCR plots. For each dataset, we plot the DCR between the holdout data and the synthetic data, and between the training data and the synthetic data for each of the three models; TVAE, CTGAN and CopulaGAN.

A.1 Dataset 1

A.1.1 Hamming

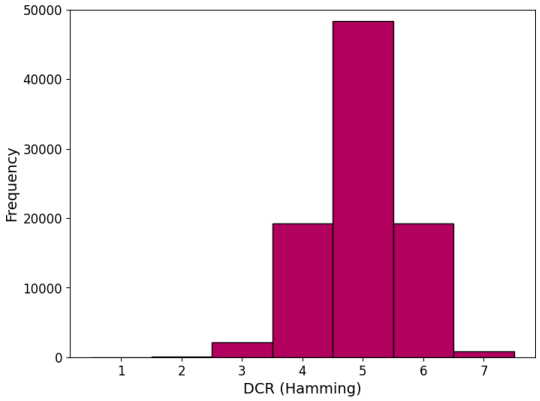
Figure 9 – DCR dataset 1 – TVAE



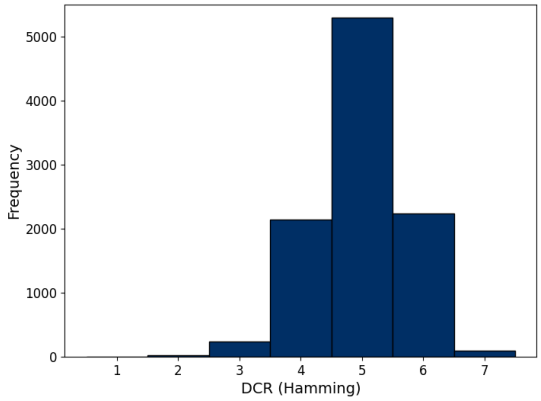
(a) Histogram of DCR between 'Training Dataset 1' and the synthetic dataset for the TVAE model based on hamming distance

(b) Histogram of DCR between 'Holdout Dataset 1' and the synthetic dataset for the TVAE model based on hamming distance

Figure 10 – DCR dataset 1 – CTGAN

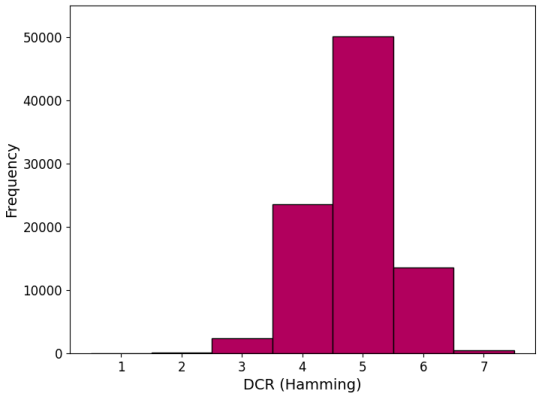


(a) Histogram of DCR between 'Training Dataset 1' and the synthetic dataset for the CTGAN model based on hamming distance

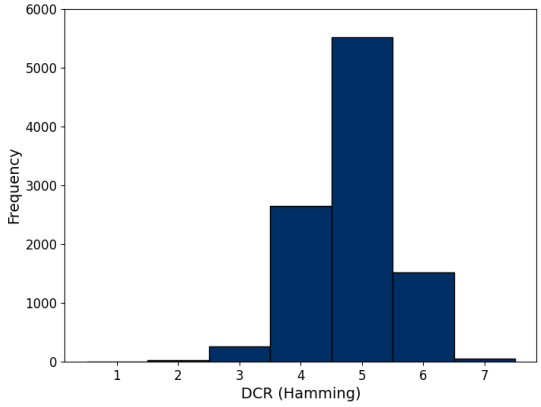


(b) Histogram of DCR between 'Holdout Dataset 1' and the synthetic dataset for the CTGAN model based on hamming distance

Figure 11 – DCR dataset 1 – CopulaGAN



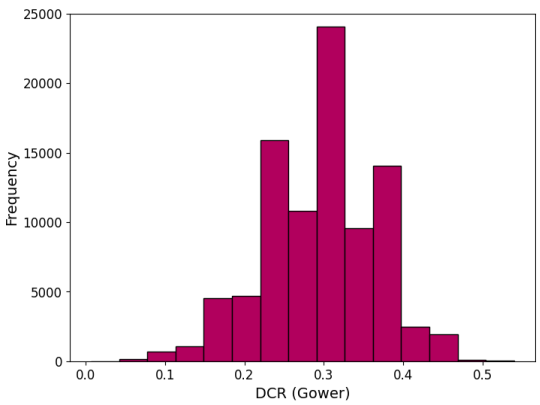
(a) Histogram of DCR between 'Training Dataset 1' and the synthetic dataset for the CopulaGAN model based on hamming distance



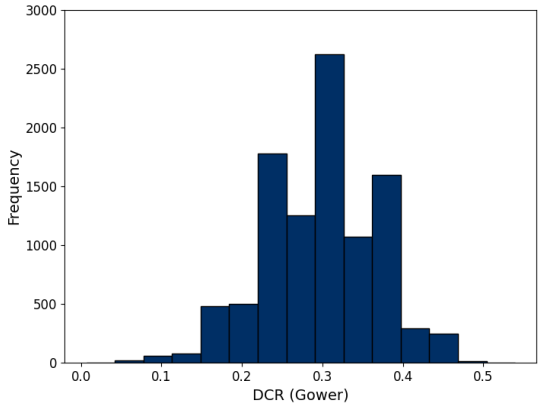
(b) Histogram of DCR between 'Holdout Dataset 1' and the synthetic dataset for the CopulaGAN model based on hamming distance

A.1.2 Gower

Figure 12 – DCR dataset 1 – TVAE

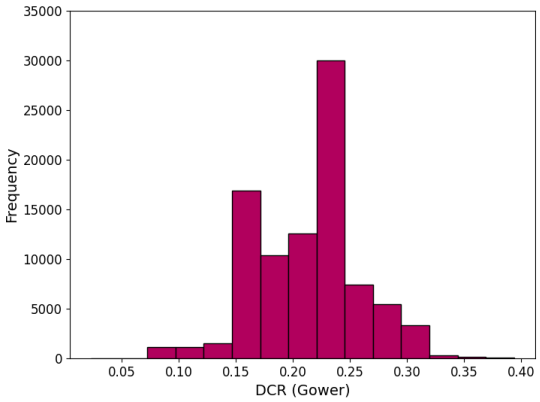


(a) Histogram of DCR between 'Training Dataset 1' and the synthetic dataset for the TVAE model based on Gower distance

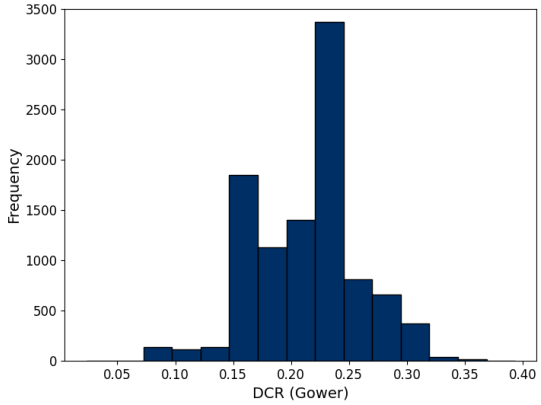


(b) Histogram of DCR between 'Holdout Dataset 1' and the synthetic dataset for the TVAE model based on Gower distance

Figure 13 – DCR dataset 1 – CTGAN

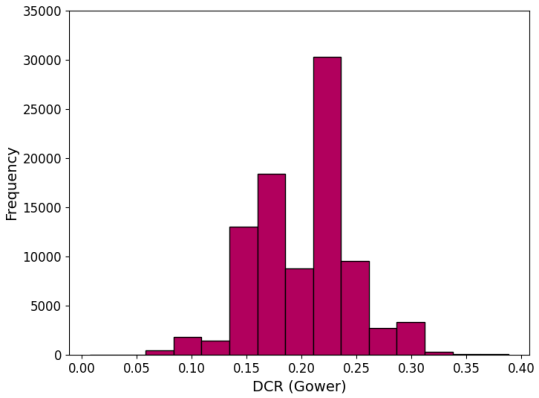


(a) Histogram of DCR between 'Training Dataset 1' and the synthetic dataset for the CTGAN model based on Gower distance

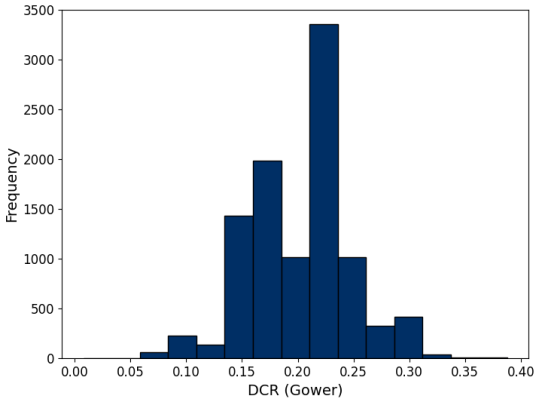


(b) Histogram of DCR between 'Holdout Dataset 1' and the synthetic dataset for the CTGAN model based on Gower distance

Figure 14 – DCR dataset 1 – CopulaGAN



(a) Histogram of DCR between 'Training Dataset 1' and the synthetic dataset for the CopulaGAN model based on Gower distance

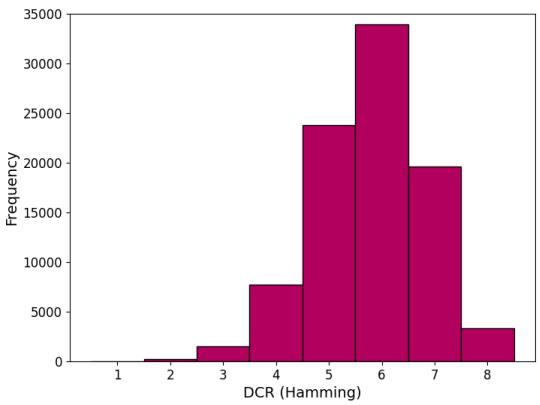


(b) Histogram of DCR between 'Holdout Dataset 1' and the synthetic dataset for the CopulaGAN model based on Gower distance

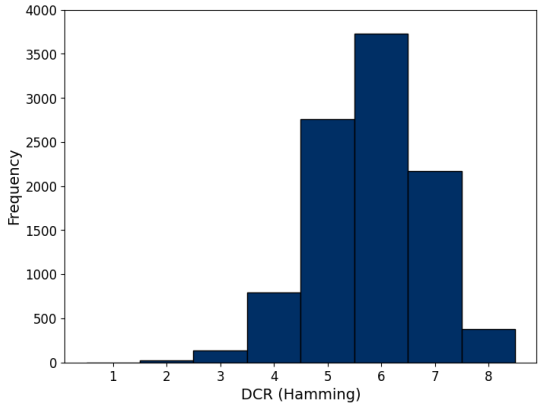
A.2 Dataset 1b

A.2.1 Hamming

Figure 15 – DCR dataset 1b - TVAE

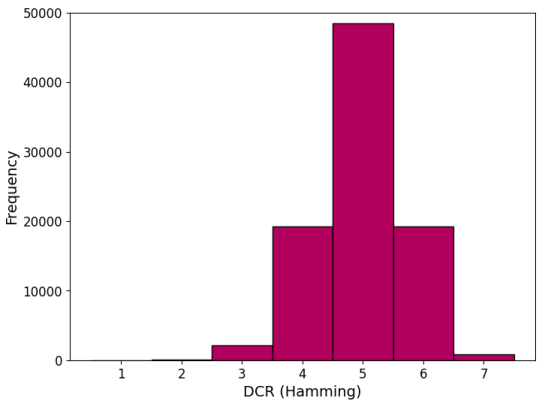


(a) Histogram of DCR between 'Training Dataset 1b' and the synthetic dataset for the TVAE model based on hamming distance

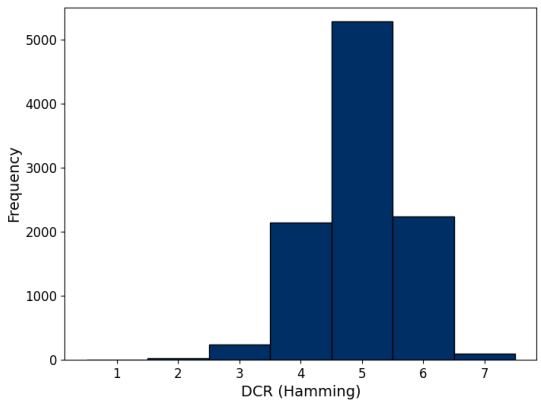


(b) Histogram of DCR between 'Holdout Dataset 1b' and the synthetic dataset for the TVAE model based on hamming distance

Figure 16 – DCR dataset 1b - CTGAN

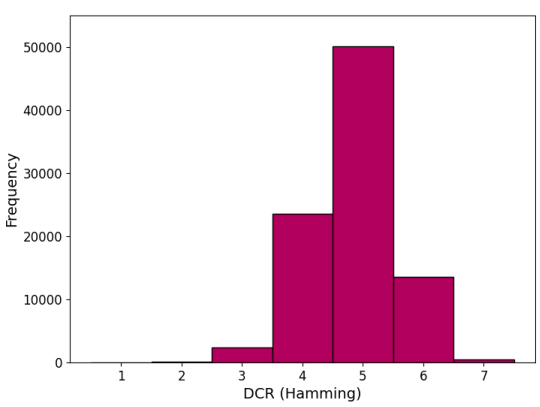


(a) Histogram of DCR between 'Training Dataset 1b' and the synthetic dataset for the CTGAN model based on hamming distance

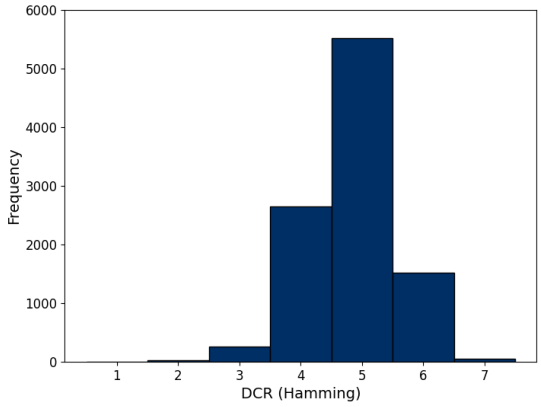


(b) Histogram of DCR between 'Holdout Dataset 1b' and the synthetic dataset for the CTGAN model based on hamming distance

Figure 17 – DCR dataset 1b - CopulaGAN



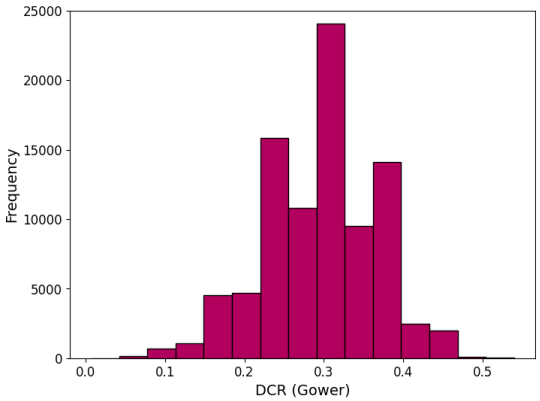
(a) Histogram of DCR between 'Training Dataset 1b' and the synthetic dataset for the CopulaGAN model based on hamming distance



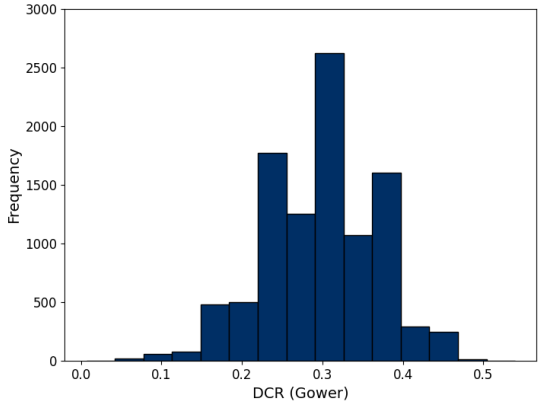
(b) Histogram of DCR between 'Holdout Dataset 1b' and the synthetic dataset for the CopulaGAN model based on hamming distance

A.2.2 Gower

Figure 18 – DCR dataset 1b – TVAE

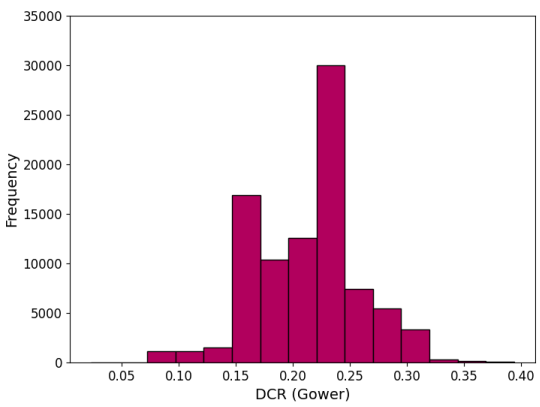


(a) Histogram of DCR between 'Training Dataset 1b' and the synthetic dataset for the TVAE model based on Gower distance

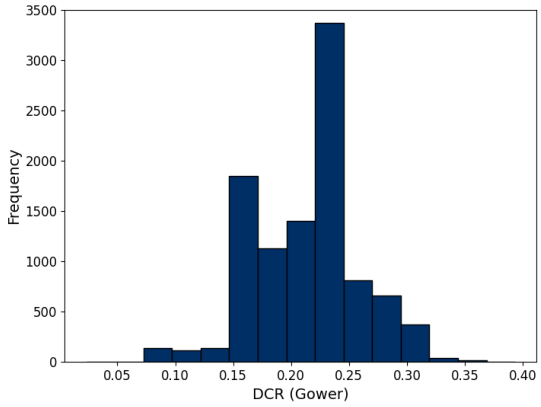


(b) Histogram of DCR between 'Holdout Dataset 1b' and the synthetic dataset for the TVAE model based on Gower distance

Figure 19 – DCR dataset 1b – CTGAN

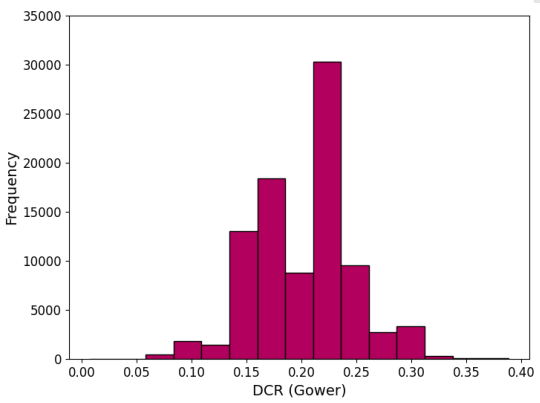


(a) Histogram of DCR between 'Training Dataset 1b' and the synthetic dataset for the CTGAN model based on Gower distance

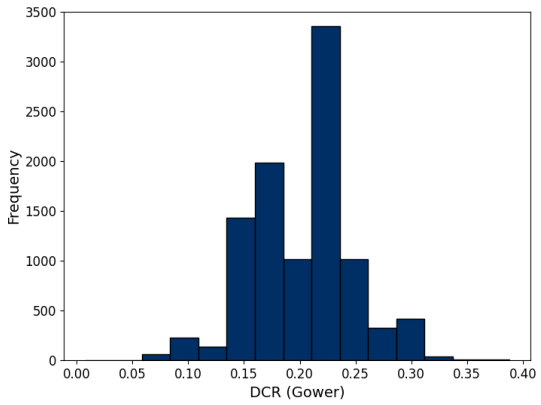


(b) Histogram of DCR between 'Holdout Dataset 1b' and the synthetic dataset for the CTGAN model based on Gower distance

Figure 20 – DCR dataset 1b – CopulaGAN



(a) Histogram of DCR between 'Training Dataset 1b' and the synthetic dataset for the CopulaGAN model based on Gower distance

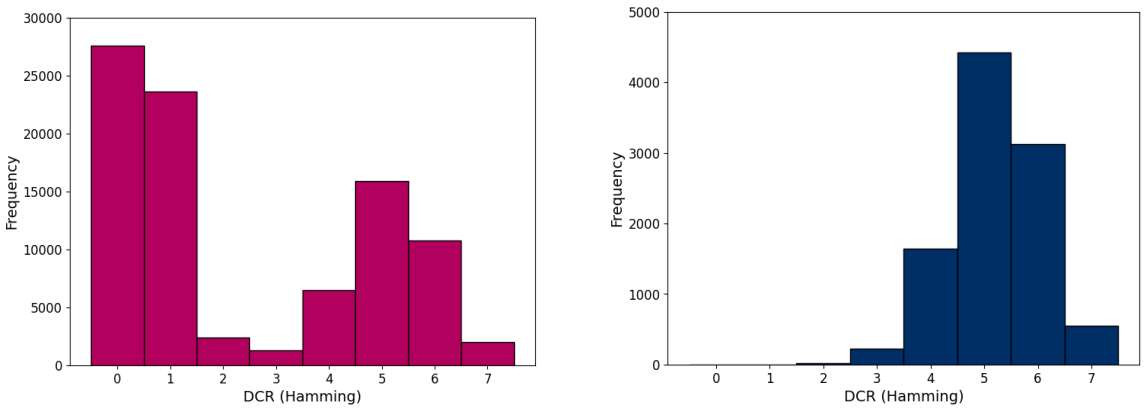


(b) Histogram of DCR between 'Holdout Dataset 1b' and the synthetic dataset for the CopulaGAN model based on Gower distance

A.3 Dataset 2

A.3.1 Hamming

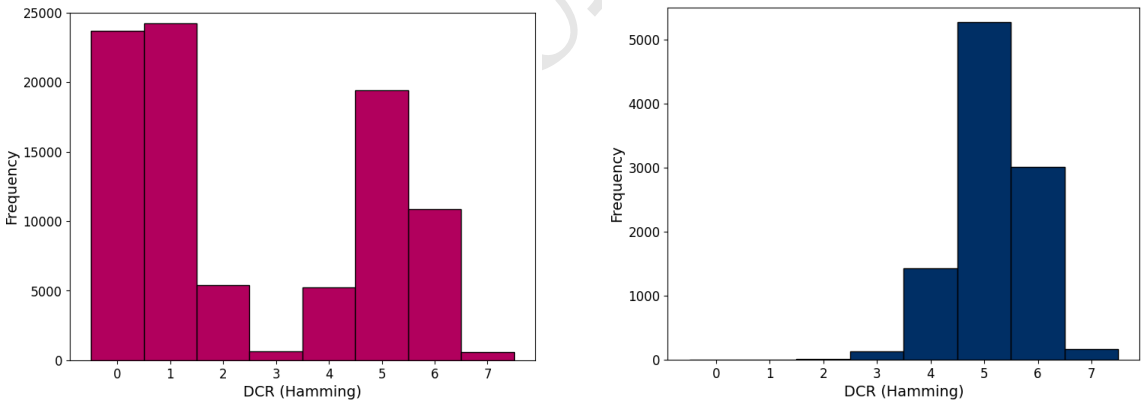
Figure 21 – DCR dataset 2 – TVAE



(a) Histogram of DCR between 'Training Dataset 2' and the synthetic dataset for the TVAE model based on hamming distance

(b) Histogram of DCR between 'Holdout Dataset 2' and the synthetic dataset for the TVAE model based on hamming distance

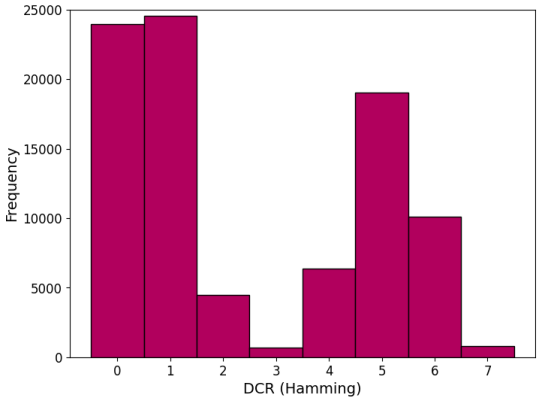
Figure 22 – DCR dataset 2 – CTGAN



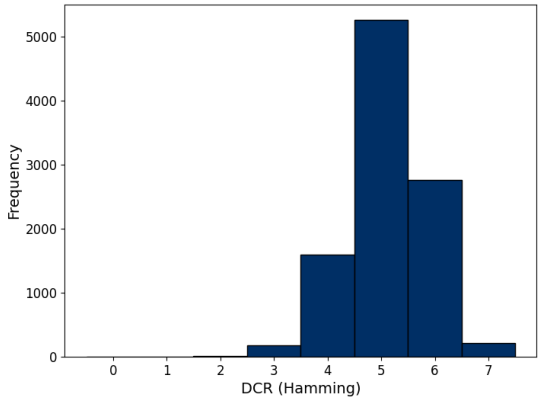
(a) Histogram of DCR between 'Training Dataset 2' and the synthetic dataset for the CTGAN model based on hamming distance

(b) Histogram of DCR between 'Holdout Dataset 2' and the synthetic dataset for the CTGAN model based on hamming distance

Figure 23 – DCR dataset 2 – CopulaGAN



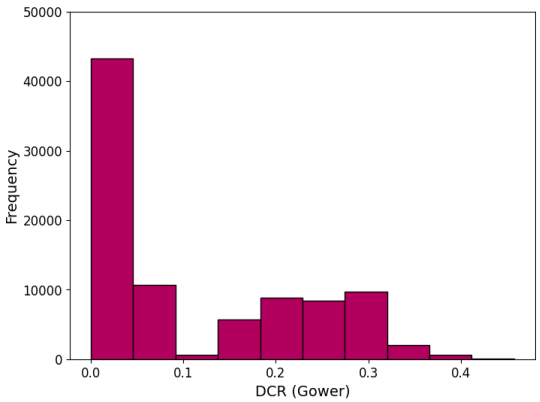
(a) Histogram of DCR between 'Training Dataset 2' and the synthetic dataset for the CopulaGAN model based on hamming distance



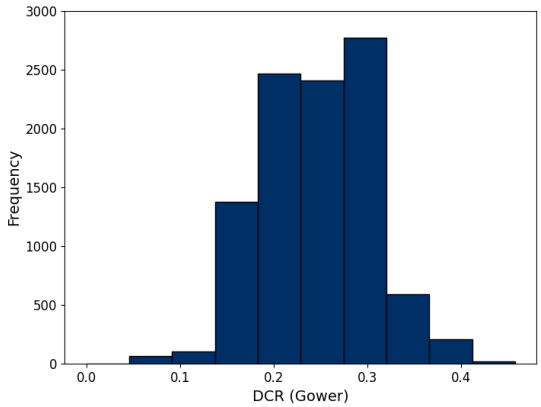
(b) Histogram of DCR between 'Holdout Dataset 2' and the synthetic dataset for the CopulaGAN model based on hamming distance

A.3.2 Gower

Figure 24 – DCR dataset 2 – TVAE

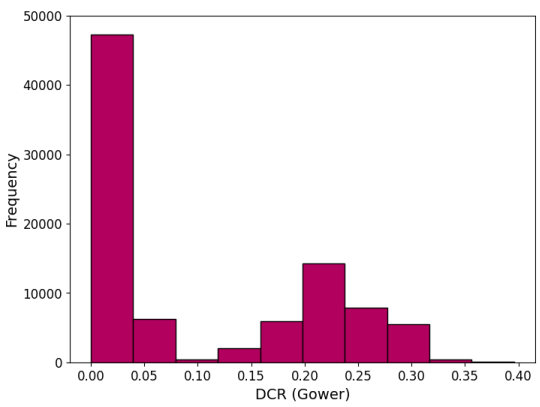


(a) Histogram of DCR between 'Training Dataset 2' and the synthetic dataset for the TVAE model based on gower distance

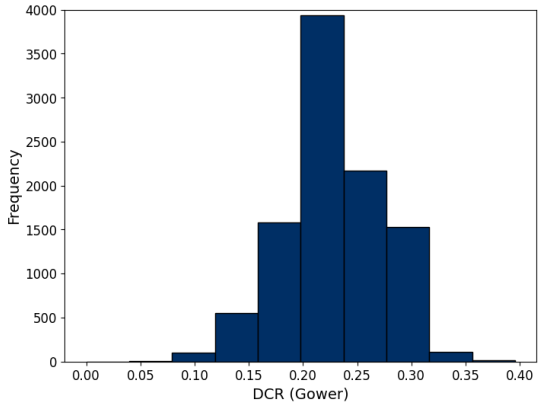


(b) Histogram of DCR between 'Holdout Dataset 2' and the synthetic dataset for the TVAE model based on gower distance

Figure 25 – DCR dataset 2 – CTGAN

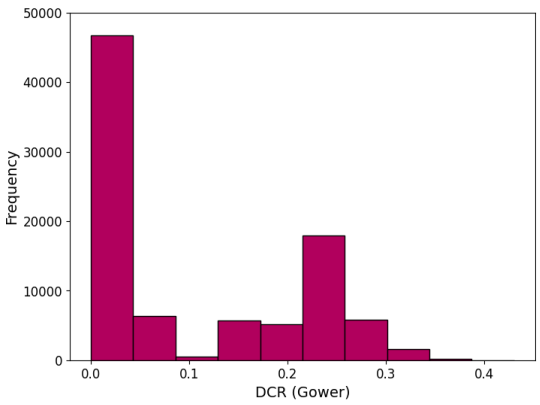


(a) Histogram of DCR between 'Training Dataset 2' and the synthetic dataset for the CTGAN model based on gower distance

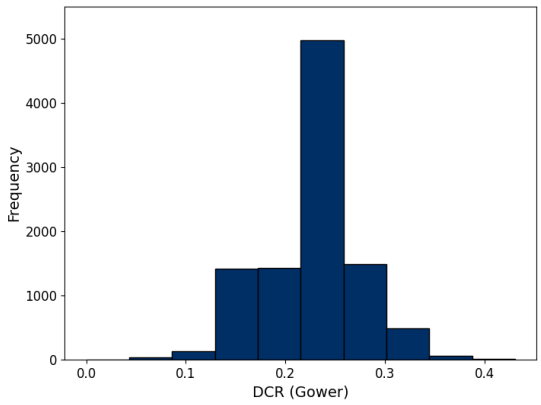


(b) Histogram of DCR between 'Holdout Dataset 2' and the synthetic dataset for the CTGAN model based on gower distance

Figure 26 – DCR dataset 2 – CopulaGAN



(a) Histogram of DCR between 'Training Dataset 2' and the synthetic dataset for the CopulaGAN model based on gower distance

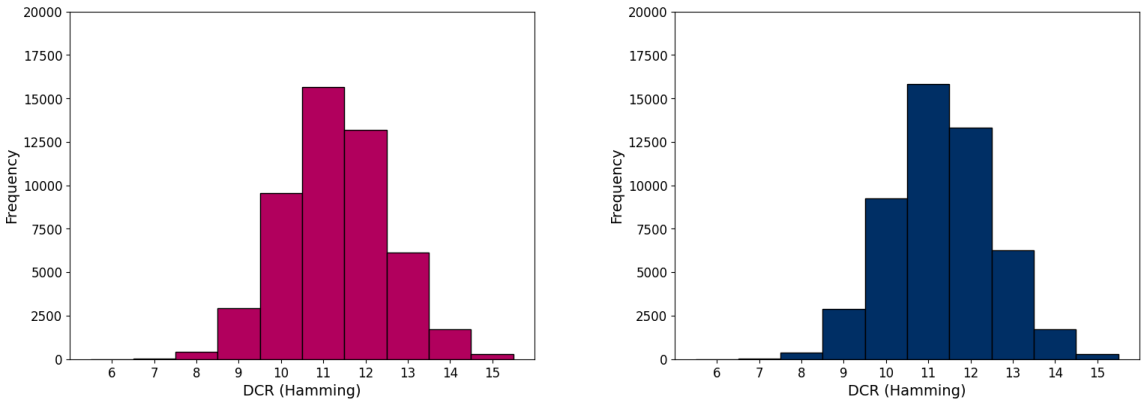


(b) Histogram of DCR between 'Holdout Dataset 2' and the synthetic dataset for the CopulaGAN model based on gower distance

A.4 Dataset 3

A.4.1 Hamming

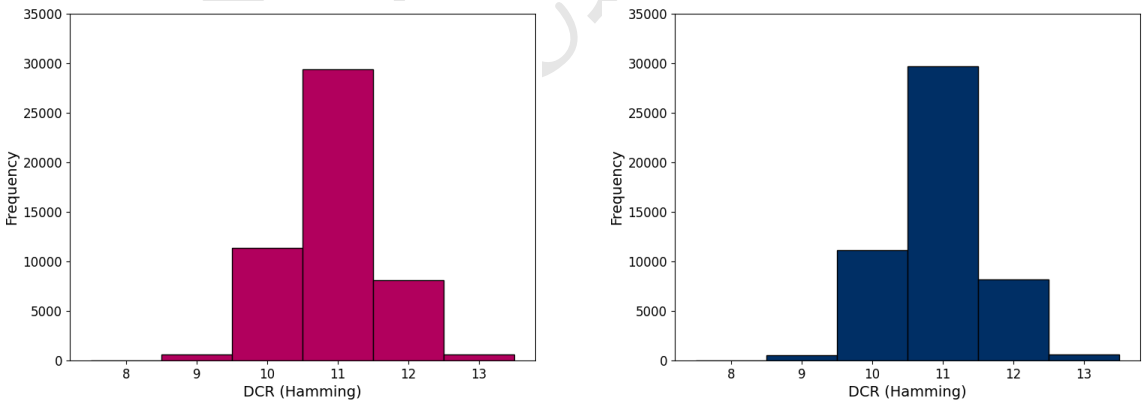
Figure 27 – DCR dataset 3 – TVAE



(a) Histogram of DCR between 'Training Dataset 3' and the synthetic dataset for the TVAE model based on hamming distance

(b) Histogram of DCR between 'Holdout Dataset 3' and the synthetic dataset for the TVAE model based on hamming distance

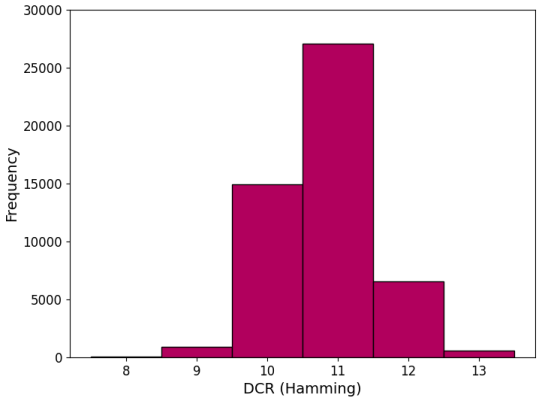
Figure 28 – DCR dataset 3 – CTGAN



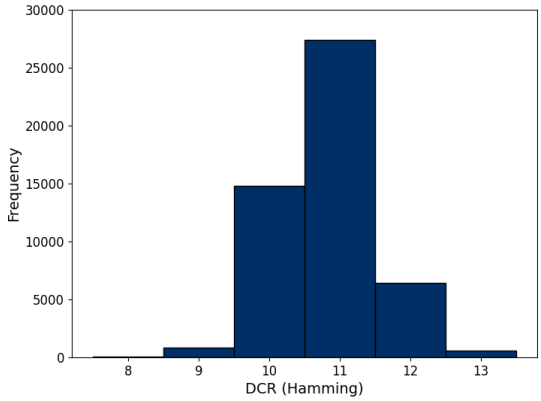
(a) Histogram of DCR between 'Training Dataset 3' and the synthetic dataset for the CTGAN model based on hamming distance

(b) Histogram of DCR between 'Holdout Dataset 3' and the synthetic dataset for the CTGAN model based on hamming distance

Figure 29 – DCR dataset 3 – CopulaGAN



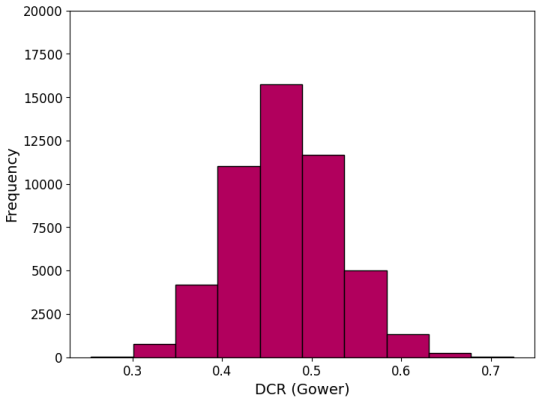
(a) Histogram of DCR between 'Training Dataset 3' and the synthetic dataset for the CopulaGAN model based on hamming distance



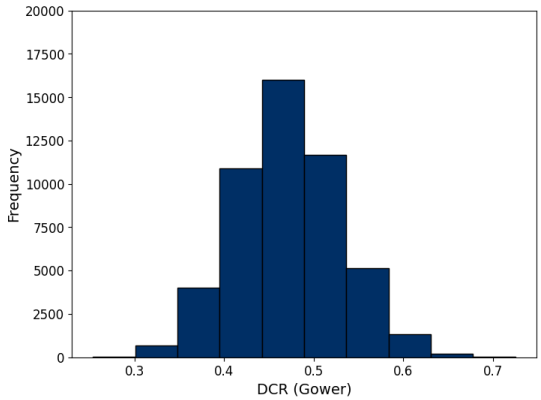
(b) Histogram of DCR between 'Holdout Dataset 3' and the synthetic dataset for the CopulaGAN model based on hamming distance

A.4.2 Gower

Figure 30 – DCR dataset 3 – TVAE

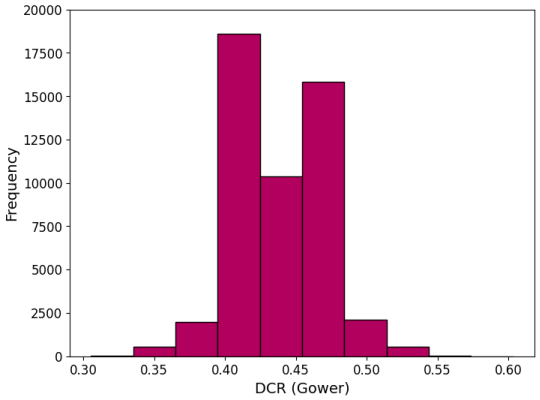


(a) Histogram of DCR between 'Training Dataset 3' and the synthetic dataset for the TVAE model based on gower distance

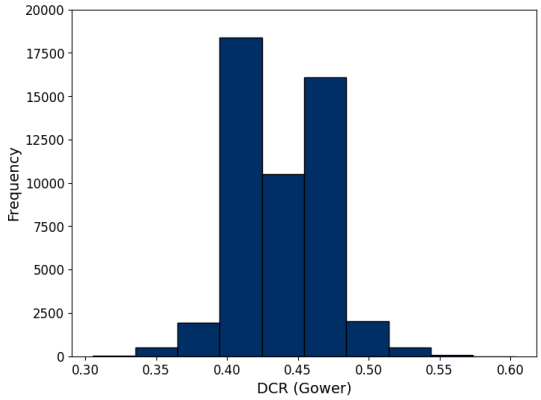


(b) Histogram of DCR between 'Holdout Dataset 3' and the synthetic dataset for the TVAE model based on gower distance

Figure 31 – DCR dataset 3 – CTGAN

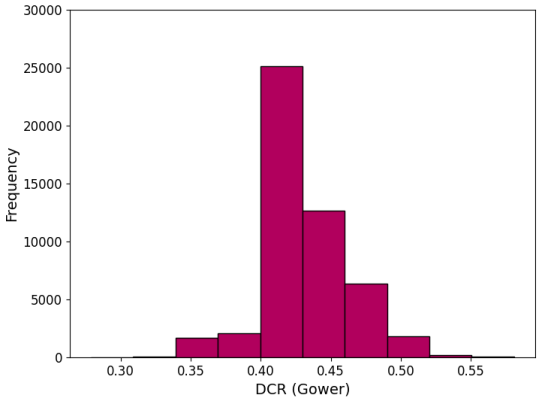


(a) Histogram of DCR between 'Training Dataset 3' and the synthetic dataset for the CTGAN model based on gower distance

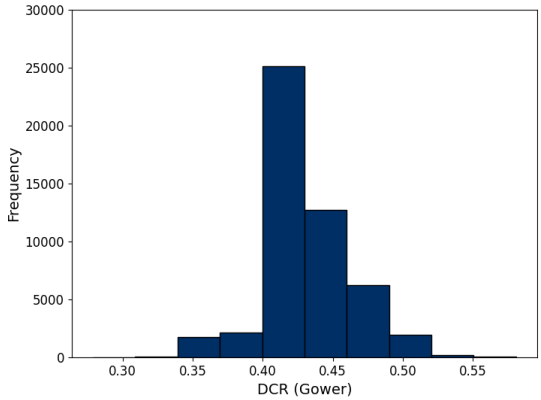


(b) Histogram of DCR between 'Holdout Dataset 3' and the synthetic dataset for the CTGAN model based on gower distance

Figure 32 – DCR dataset 3 – CopulaGAN



(a) Histogram of DCR between 'Training Dataset 3' and the synthetic dataset for the CopulaGAN model based on gower distance

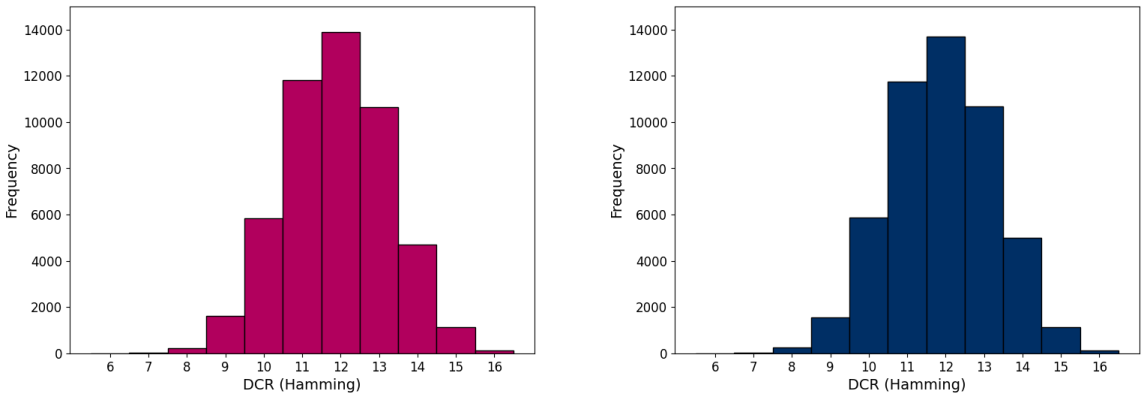


(b) Histogram of DCR between 'Holdout Dataset 3' and the synthetic dataset for the CopulaGAN model based on gower distance

A.5 Dataset 4

A.5.1 Hamming

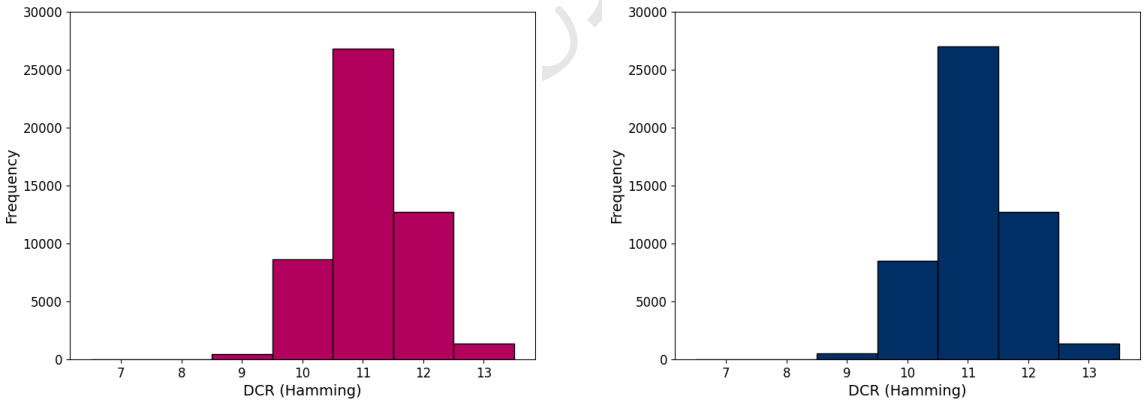
Figure 33 – DCR dataset 4 – TVAE



(a) Histogram of DCR between 'Training Dataset 4' and the synthetic dataset for the TVAE model based on hamming distance

(b) Histogram of DCR between 'Holdout Dataset 4' and the synthetic dataset for the TVAE model based on hamming distance

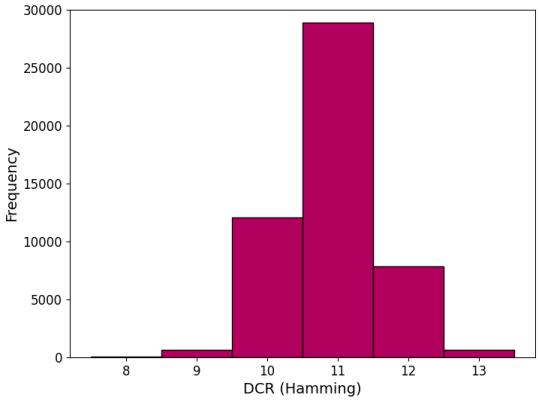
Figure 34 – DCR dataset 4 – CTGAN



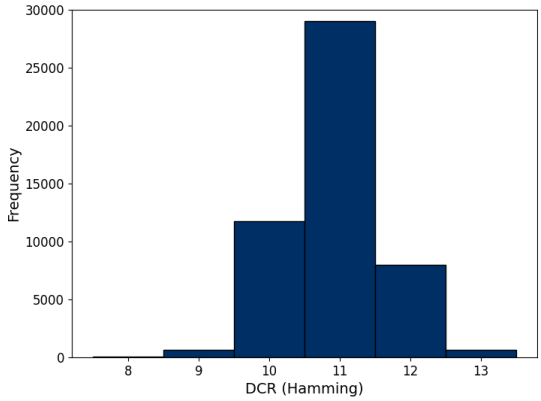
(a) Histogram of DCR between 'Training Dataset 4' and the synthetic dataset for the CTGAN model based on hamming distance

(b) Histogram of DCR between 'Holdout Dataset 4' and the synthetic dataset for the CTGAN model based on hamming distance

Figure 35 – DCR dataset 4 – CopulaGAN



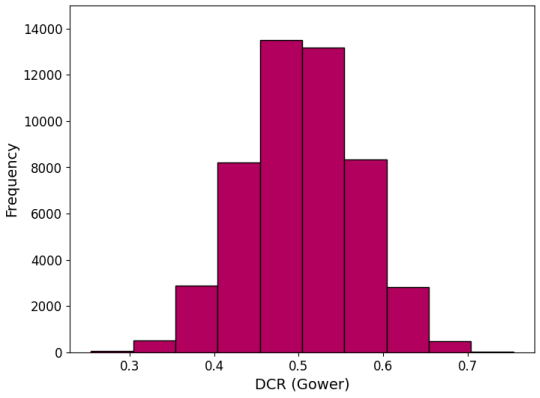
(a) Histogram of DCR between 'Training Dataset 4' and the synthetic dataset for the CopulaGAN model based on hamming distance



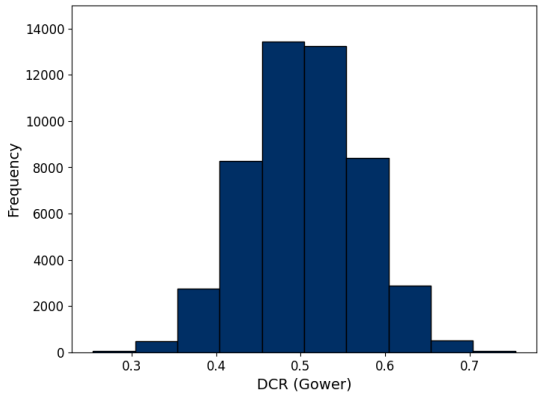
(b) Histogram of DCR between 'Holdout Dataset 4' and the synthetic dataset for the CopulaGAN model based on hamming distance

A.5.2 Gower

Figure 36 – DCR dataset 4 – TVAE

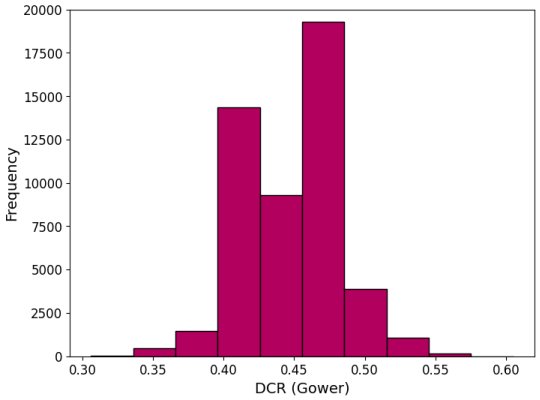


(a) Histogram of DCR between 'Training Dataset 4' and the synthetic dataset for the TVAE model based on gower distance

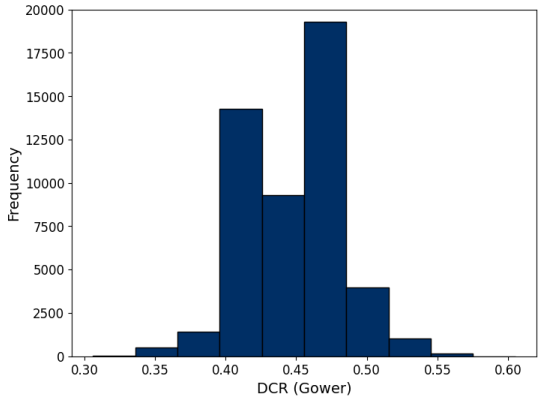


(b) Histogram of DCR between 'Holdout Dataset 4' and the synthetic dataset for the TVAE model based on gower distance

Figure 37 – DCR dataset 4 – CTGAN

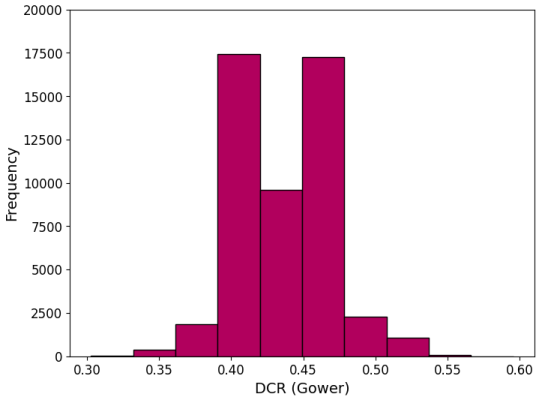


(a) Histogram of DCR between 'Training Dataset 4' and the synthetic dataset for the CTGAN model based on gower distance

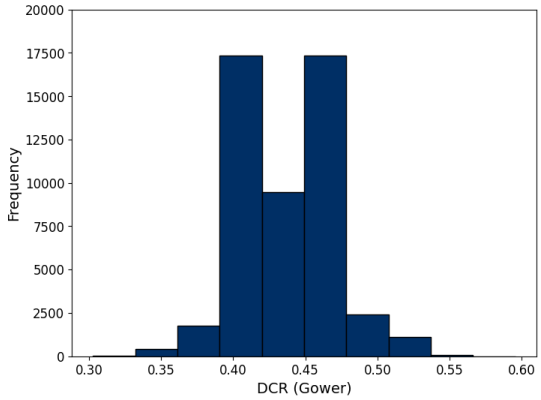


(b) Histogram of DCR between 'Holdout Dataset 4' and the synthetic dataset for the CTGAN model based on gower distance

Figure 38 – DCR dataset 4 – CopulaGAN



(a) Histogram of DCR between 'Training Dataset 4' and the synthetic dataset for the CopulaGAN model based on gower distance

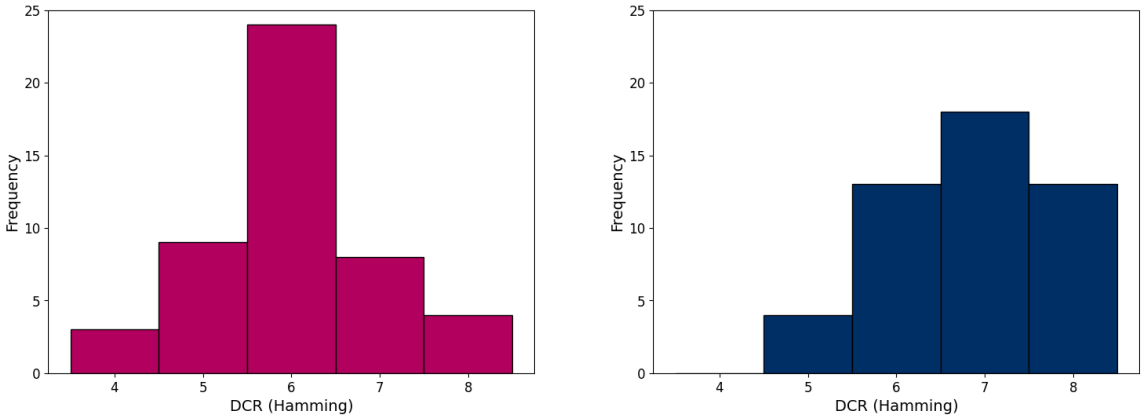


(b) Histogram of DCR between 'Holdout Dataset 4' and the synthetic dataset for the CopulaGAN model based on gower distance

A.6 Dataset 5

A.6.1 Hamming

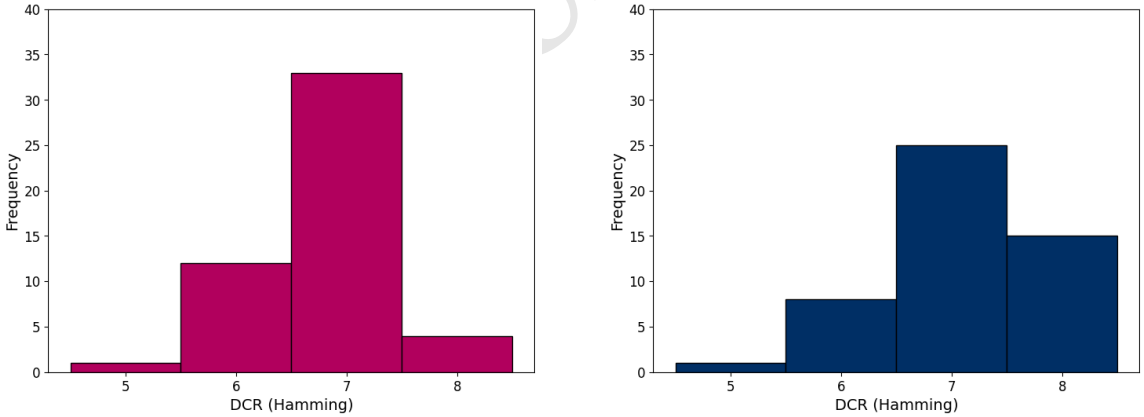
Figure 39 – DCR dataset 5 – TVAE



(a) Histogram of DCR between 'Training Dataset 5' and the synthetic dataset for the TVAE model based on hamming distance

(b) Histogram of DCR between 'Holdout Dataset 5' and the synthetic dataset for the TVAE model based on hamming distance

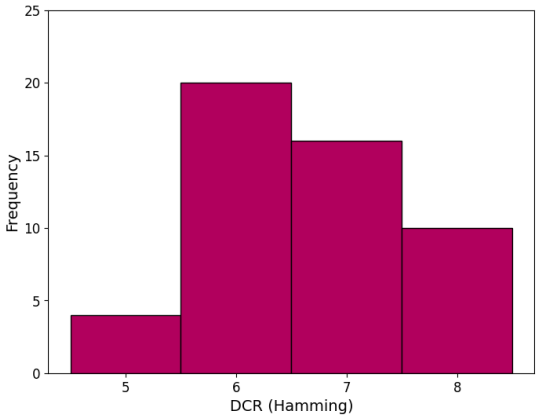
Figure 40 – DCR dataset 5 – CTGAN



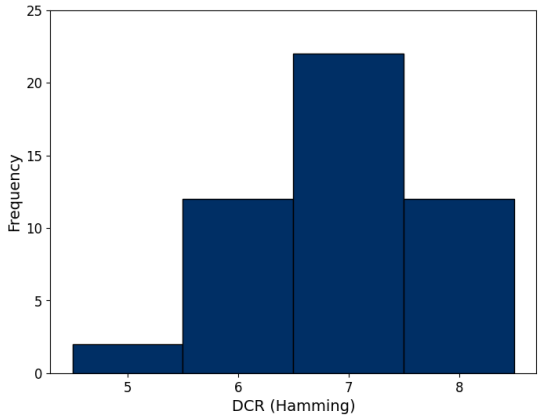
(a) Histogram of DCR between 'Training Dataset 5' and the synthetic dataset for the CTGAN model based on hamming distance

(b) Histogram of DCR between 'Holdout Dataset 5' and the synthetic dataset for the CTGAN model based on hamming distance

Figure 41 – DCR dataset 5 – CopulaGAN



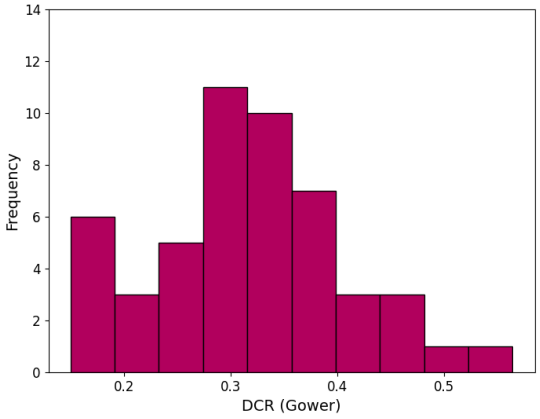
(a) Histogram of DCR between 'Training Dataset 5' and the synthetic dataset for the CopulaGAN model based on hamming distance



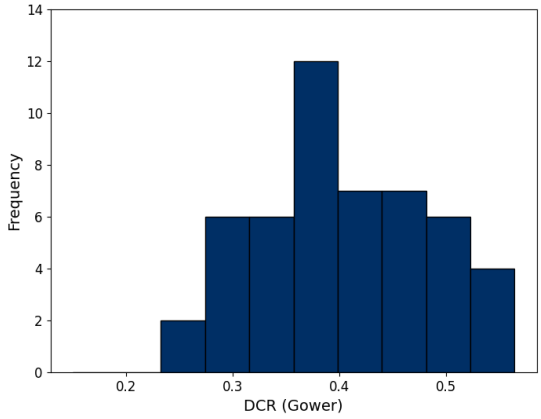
(b) Histogram of DCR between 'Holdout Dataset 5' and the synthetic dataset for the CopulaGAN model based on hamming distance

A.6.2 Gower

Figure 42 – DCR dataset 5 – TVAE

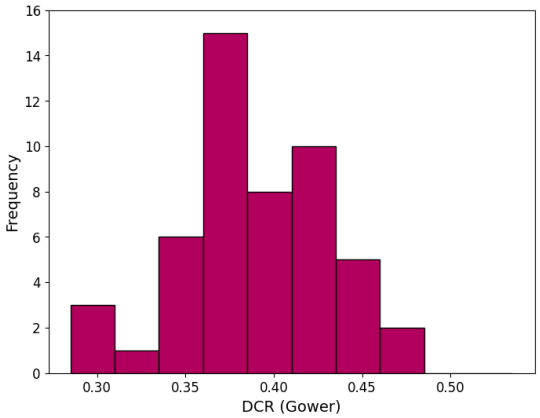


(a) Histogram of DCR between 'Training Dataset 5' and the synthetic dataset for the TVAE model based on gower distance

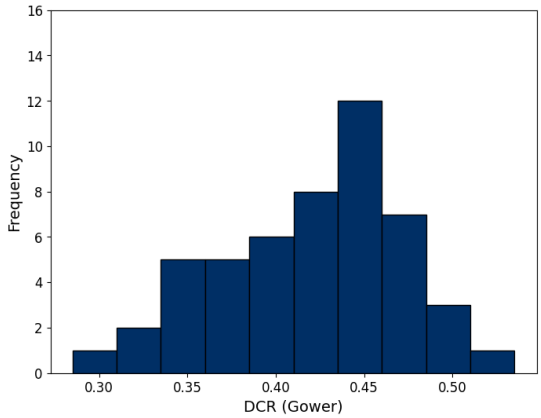


(b) Histogram of DCR between 'Holdout Dataset 5' and the synthetic dataset for the TVAE model based on gower distance

Figure 43 – DCR dataset 5 – CTGAN

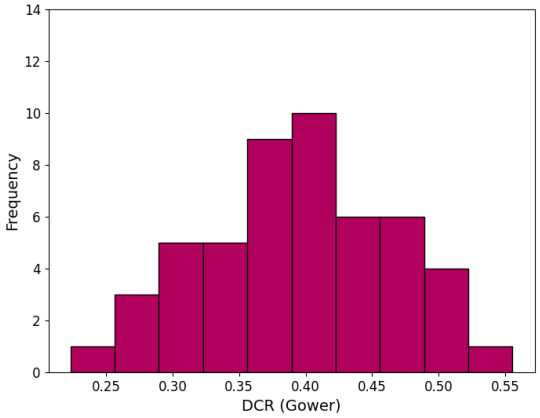


(a) Histogram of DCR between 'Training Dataset 5' and the synthetic dataset for the CTGAN model based on gower distance

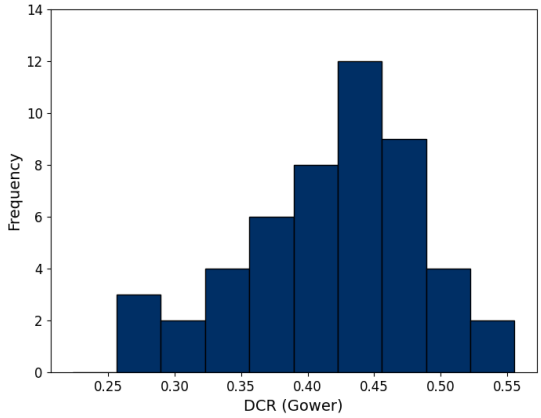


(b) Histogram of DCR between 'Holdout Dataset 5' and the synthetic dataset for the CTGAN model based on gower distance

Figure 44 – DCR dataset 5 – CopulaGAN



(a) Histogram of DCR between 'Training Dataset 5' and the synthetic dataset for the CopulaGAN model based on gower distance



(b) Histogram of DCR between 'Holdout Dataset 5' and the synthetic dataset for the CopulaGAN model based on gower distance