Minh Vu

QUALITY ASSESSMENT OF SYNTHETIC HEALTH DATA

Bachelor of Science Thesis
Faculty of Medicine and Health Technology
Examiners: Mark van Gils, Saana Seppälä
April 2025

ABSTRACT

Minh Vu: Quality Assessment of Synthetic Health Data
Bachelor of Science Thesis
Tampere University
Bachelor's Degree Program in Science and Engineering
April 2025

Synthetic health data is artificially generated to replicate the characteristics of real-world health data without exposing sensitive patient information. Its popularity has surged in recent years due to its potential to replace real health data in research while safeguarding patient privacy. However, ensuring its quality requires a systematic, logical, and thorough evaluation. Since synthetic health data can potentially be used in research areas such as disease analysis and medical device testing, rigorous quality assessment is essential, particularly in healthcare, where every action must be carefully justified to protect patient safety. Currently, research on evaluating synthetic health data is limited, especially for non-image-based data. Moreover, existing studies vary significantly, focusing on different aspects of data quality without a unified terminology across the scientific community.

This study aims to comprehensively present the quality metrics of synthetic health data and the methods used to assess them, focusing on synthetic electronic health records and biomedical signals. To achieve this, a literature review was conducted on 54 scientific articles that were carefully selected to ensure both their relevance and incorporation of state-of-the-art insights. Information from these articles was properly integrated to achieve this study's outcomes. As a result, five key dimensions of synthetic health data are identified, with three explored in greater detail through corresponding evaluation methods and metrics. Additionally, an evaluation framework is introduced to demonstrate how these quality dimensions can be assessed in practice.

Various methods and metrics exist for assessing quality dimensions, yet they also have limitations, such as subjective judgment or indifference to the temporal nature of synthetic health data. There is significant potential for advancing healthcare research through synthetic health data. Thus, to maximize its impact and enable widespread real-world applications, future studies are needed to address these limitations.

Keywords: Synthetic health data, Quality assessment, Electronic health records, Biomedical signals, Health informatics.

USE OF AI IN THE THESIS

I have utilised AI tools in my thesis:

☐ No
☒ Yes

The AI tools utilised in my thesis and their purposes are described below:

Names and versions of AI tools: ChatGPT (March 31, 2025 version).

Purpose of using AI tools and the parts in which they were used: This AI tool was used with two main purposes. The first is to extract useful information from the selected articles during the literature review. ChatGPT provides an effective way to summarize long and intricate pieces of text, so using it for this purpose made the whole thesis process less laborious. Secondly, ChatGPT was used to increase the readability of the text that I originally produced. ChatGPT rephrased my original text so that readers would have the most pleasant time reading my thesis, as my ideas are presented in the easiest way to understand. My goal when using ChatGPT for this purpose is that even readers who are not majoring in Biomedical Engineering or Data Sciences can still understand the basics of my thesis. Thus, no ideas presented in this thesis are generated by ChatGPT; rather, several of my ideas are presented in ChatGPT's words. ChatGPT was used in the second paragraph of Section 2.1, the first paragraph of Section 3.1, the second paragraph of Section 4.2.1, Section 4.2.4, the second paragraph of Chapter 5, the first paragraph of Section 5.4, Section 6.2, and the second paragraph of Chapter 7.

I am aware that I am totally responsible for the entire content of the thesis, including the parts generated by AI, and accept the responsibility for any violations of the ethical standards of publications.

PREFACE

I would like to express my deep gratitude to the two supervisors of this thesis, Professor Mark van Gils and Doctoral Researcher Saana Seppälä, for their dedication in guiding me towards the completion of my thesis. Additionally, I would like to thank my family and close friends for their unwavering moral support. This thesis is an important milestone in my academic journey.

Tampere, April 20, 2025

Minh Vu

List of Symbols and Abbreviations

AA	Adversarial accuracy
AI	Artificial intelligence
AIA	Attribute Inference Attack
CCD	Cross-type conditional distribution
CVT	Constraint Violation Test
CS	Cosine similarity
CTGANs	Conditional Tabular GAN
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DM	Diffusion Model
DP	Differential Privacy
DWP	Dimension-wise prediction
ECG	Electrocardiogram
EHR	Electronic health record
FOP	First-order proximity
GAN	Generative Adversarial Network
JSD	Jensen Shannon divergence
KLD	Kullback–Leibler divergence
MAE	Mean Absolute Error
MDS	Multidimensional Scaling
MIA	Membership Inference Attack
ML	Machine Learning
MSE	Mean Squared Error
PCA	Principal Component Analysis
PCD	Pairwise Correlation Difference
SD	Synthetic data
SDG	Synthetic data generation
SHD	Synthetic health data
t-SNE	t-distributed Stochastic Neighbor Embedding
TRTR	Train on Real, Test on Real
TRTS	Train on Real, Test on Synthetic
TSTR	Train on Synthetic, Test on Real
UMAP	Uniform Manifold Approximation and Projection
VAE	Variational Autoencoder

CONTENTS

1. INTRODUCTION

Information and Communication Technologies have impacted all aspects of life, revolutionizing hospitals from human-centric practices to technology-assisted care. Its integration in medicine is evident in telemedicine, Electronic Health Records (EHRs) in complex medical informatics, personal health assistants, and decision support systems, significantly improving efficiency, accuracy, and reliability. Rapid technological advances drive the constant adoption of new systems, which rely on high-quality healthcare data for research, development, and decision-making. Ensuring the quality of digital health systems requires large datasets for clinical validation. Regrettably, accessing healthcare data presents unique challenges, mainly due to privacy concerns. Medical datasets often contain sensitive personal information, such as diagnoses, treatments, and billing records, making open sharing undesirable and raising ethical, financial, and legal issues. Researchers have proposed solutions like data masking and anonymization. However, synthetic data (SD) stands out as the most effective solution. (1)

Synthetic data is artificially created data that retains the characteristics of the original data while not containing the personal information of actual patients. Research on synthetic medical data generation has surged in recent years, with broader adoption anticipated in the future. Researchers are exploring various methods to create realistic synthetic data and have introduced multiple quality evaluation metrics to assess its viability as a substitute for real data. (1) Nevertheless, they prioritized different aspects of quality due to the lack of a unified methodology for synthetic data quality assessment. Thus, this thesis seeks to identify and consolidate various quality metrics for evaluating synthetic data, along with the methods that utilize these metrics. This was achieved through a literature review, integrating insights from multiple sources to present the methods and metrics clearly and comprehensively. Notably, research on synthetic medical images is covered only briefly due to its extensive presence in the scientific community.

This chapter is followed by Chapter 2, where background information on synthetic data is presented. Chapter 3 outlines the methodology used for the literature review. Chapter 4 introduces four quality dimensions of synthetic data, along with the corresponding available methods and metrics for evaluation. Chapter 5 demonstrates how these quality dimensions were measured experimentally and proposes a systematic quality evaluation framework. Chapter 6 discusses key limitations in methods and metrics found in Chapter 4, while Chapter 7 presents key remarks on this thesis. Finally, Chapter 8 concludes this study.

2. THEORETICAL BACKGROUND

To facilitate readers' understanding of the core research problems and the results presented in this thesis, this chapter starts with a comprehensive introduction to synthetic health data. After that, the significance of evaluating synthetic health data will be covered, and this chapter will end with a brief discussion of Machine Learning (ML) and the main generative methods employed to synthesize data.

2.1 Introduction to Synthetic Health Data (SHD)

Synthetic data (SD) refers to artificially generated datasets that aim to replicate real-world data's statistical and structural characteristics (2,3). Various introductions to synthetic data are presented in selected articles, but the recently suggested definition from The Royal Society and The Alan Turing Institute is concise and comprehensive, capturing the essence of synthetic data. According to them, synthetic data is "data created using a specifically designed mathematical model or algorithm to address one or more data science tasks" (4).

Particularly, synthetic data strives to retain the utility of the original data for analysis or algorithm development while closely mimicking real-world datasets' distributions, dependencies, and patterns (3,5,6). Moreover, it ensures that individual records are untraceable, mitigating privacy risks as privacy regulations are strict in healthcare (2,3,5). Techniques like generative models (e.g., Generative Adversarial Networks, autoencoders), rule-based systems, and probabilistic algorithms are commonly used to create synthetic datasets tailored to specific applications or domains (2,3,5). Synthetic data can be used in various applications, including research, education, evaluation of methodologies (such as data linkage or machine learning algorithms), and addressing privacy concerns (1,2,4).

Recently, deep generative models have become increasingly popular across many fields for creating synthetic data, and healthcare is no exception (7). Generally speaking, synthetic health data (SHD) is used to replace real health data in certain phases of the research and development phases, and the main reason for this practice is privacy concerns (from this point onward, SD refers to SHD). For example, personal information is usually contained in Electronic Health Records (EHRs), delaying their usage in secondary research (due to regulations and ethical considerations), such as building machine learning models or testing new medical devices (1,5,8). As a solution, creating synthetic data can reduce the risk of re-identification, enhance patient privacy, and streamline data utility agreements, as they are, ideally, untraceable to real persons (4). For instance, a synthetic data generator could create a vast dataset, enabling extensive research into new models that help understand the etiology of Myalgic Encephalomyelitis or chronic fatigue syndrome (9). Besides addressing privacy issues, synthetic data is popularly used in augmenting datasets, particularly rebalancing datasets by generating synthetic records for underrepresented groups (5). Additional scenarios where synthetic health data has been employed are first-order feasibility, where institutions use synthetic data to cheaply estimate the potential value of real data for a specific task before investing in its procurement, data science competitions and hackathons, and academic settings, where synthetic data is used for a classroom data science challenge (1).

2.2 Significance of Evaluating Synthetic Health Data

Synthetic data is still an emerging approach in healthcare; thus, there are no unified metrics for assessing synthetic data's generic quality yet (1). Multiple articles share the same point of view: There is a critical need for a standardized approach to evaluate the "goodness" of synthetic data, especially in a sensitive field like healthcare, where every action must be thoroughly validated to ensure the patient's safety. According to Vallevik et al. (10), there are different evaluation frameworks where different quality dimensions are prioritized, and the terminology used in those researches was disparate. As an example, a study in 2023 by Hernandez, Epelde, Alberdi, Cilla, and Rankin examined three quality dimensions: Resemblance, Utility, and Privacy. On the other hand, a 2022 research by Yan et al. studied two main groups of quality dimensions: Data Utility, which was further divided into Resemblance and Outcome prediction, and Privacy. Furthermore, El Emam considered Utility and "Bias and stability assessment" as two primary dimensions of quality. This reality highlights the need for a standardized approach Further than that, a standardized evaluation of SD's quality allows for consistent comparison of different generative models and builds trust in SD among stakeholders, such as healthcare professionals and researchers, who may otherwise be hesitant to use synthetic data due to concerns about accuracy and reliability (1,4).

2.3 Machine Learning and Generative Methods

This subchapter begins with a concise and simple introduction to Machine Learning (ML) and its sub-concepts, providing a solid foundation for understanding the theory presented in the following paragraphs. Artificial intelligence (AI) refers to computer systems that mimic human learning, reasoning, and problem-solving. Machine learning (ML), a subset of AI, uses vast data and computing power to identify patterns and train models without explicit programming. ML can be classified into supervised, unsupervised, and reinforcement learning, and it continuously improves its performance by analyzing and adapting to new training data. Deep learning, a specialized form of ML, uses multiple layers of abstraction to extract complex patterns from data, enabling the learning of intricate functions through both supervised and unsupervised algorithms (11). Generative modeling is an ML approach where models are trained to create new data resembling a given dataset. These models are typically probabilistic, incorporating randomness to generate diverse outputs, rather than purely deterministic, which would produce fixed, identical results (12). Deep generative models can be understood as generative models developed using deep learning architectures.

Synthetic Data Generation (SDG) fundamentally involves creating a model that captures the “ground truth” in real data (1). Within the healthcare sector, Synthetic Data Generation (SDG) has been explored for various data natures. According to Hernadez M et al., biomedical signals, medical images, electronic health record (EHR) free-text notes, time-series data from smart homes and living labs activity data, and tabular data extracted from EHRs are the main forms of synthetic health data that have been examined (13). In this thesis work, SHD is categorized into biomedical signals, medical images, and EHRs in tabular format. The majority of records utilized in this literature review emphasize the use of deep generative models, with three sub-classes: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models (DMs), as the approach for data synthesizing. The selected records show that these three modern machine learning models have been studied with all three forms of SHD, indicating their dominance in the field of health data generation. Additionally, techniques like Classification and Regression Trees (CARTs), classic models like Copulas, Monte Carlo, and Bayesian Networks, and platforms such as MDClone are notable methods out of several approaches appropriate for synthesizing data. Owing to the prevalence and sophistication of deep generative models and the emphasis on the privacy preservation of MDClone, the following subsections will review GANs, VAEs, DMs, and the MDClone platform more comprehensively.

2.3.1 Generative Adversarial Networks (GANs)

SDG methods are broadly grouped into three main categories based on the source of ground truth: Knowledge-Driven, Data-Driven, and Hybrid (1). Data-driven approaches develop the generative model from real data. They have been examined more meticulously than theory-based techniques primarily due to the ability of real field data to capture ground realities. Furthermore, Data-driven methods are efficient because they can be autonomous. Data-driven methods can be further categorized into transformative and simulation modes. Simulation techniques involve generative models that learn the underlying probability distribution of real data, and synthetic samples are then generated by drawing from this learned distribution. A GAN is a typical representative of a simulation technique. (1)

GANs are a class of deep learning architectures comprising two neural networks: a Generator and a Discriminator, and these networks function as adversaries in a competitive process. GANs are utterly designed to generate synthetic data indistinguishable from real data. The input to the Generator is noise, while its output is synthetic data. The Discriminator has two inputs: the real training data and the synthetic data generated by the Generator, and its output indicates whether its input is real or synthetic. The Generator is trained to produce SD that is as close as possible to real data to fool the Discriminator, while the Discriminator aims to maximize its discriminatory capability (14,15). Finally, when the generated samples follow the same distribution as the real data, the Discriminator can only guess randomly (16). Figure 1 summarizes this introduction to GANs by providing a simple visual representation of the essence of GANs in the context of synthesizing Electrocardiogram (ECG) signals.

***Figure 1**. A general design of GAN. GAN consists of the Generator, which takes noise as input and aims to synthesize realistic SD, and the Discriminator, which takes real and fake data as input and aims to indicate whether its input is real or synthetic. (15)*

Multiple specific models of the GANs family have been developed to serve the developers' demands. Recurrent GANs (RGANs) and Recurrent Conditional GANs (RCGANs) have been researched to produce real-valued, multidimensional time-series medical data, followed by the exploitation of RGANs to synthesize ECG signals (15,17). Conditional GANs (CGANs) have been used with static, time-invariant EHRs. The real-world time-series GANs (RTSGANs) can be employed to synthesize medical data with irregular and variable temporal patterns, while TimeGANs are also able to create time-series EHR (17,18). HealthGANs and Conditional Tabular GANs (CTGANs) can effectively generate SD with numerical, binary, and categorical features (13). The family of GANs is prolific, and since the main focus of this thesis is not SD generation techniques, the coverage of GANs in this work is limited. Readers interested in this topic can refer to (1) to gain a deeper understanding of contemporary GAN architectures.

2.3.2 Variational Autoencoders (VAEs)

Autoencoders (AEs) are neural networks consisting of the Encoder, which compresses the input data into lower dimensionality in the latent space, and the Decoder, which aims to reconstruct the original data from its encoded version (14,15). AEs strive to learn the dimensional dependencies within the input data to generate new samples, and they can work as standalone generative models or within GAN-based architectures (1,17). E. Budu et al. characterized VAEs as AEs with a probabilistic nature (17). In VAEs, the Encoder acquires latent variables and their posteriors from the original data. Next, values are extracted from the latent space and used as input for the Decoder. The Decoder's function is to recreate the original data by learning the mapping from the latent space to the input space (17). In synthesizing the ECG signal, the input signal is encoded into a distribution of variables, allowing the Decoder to reconstruct new signals by sampling from this distribution and generating new ECGs (15). Figure 2 offers a concise visual summary of this introduction to VAEs, illustrating the core concept of VAEs in generating ECG signals.

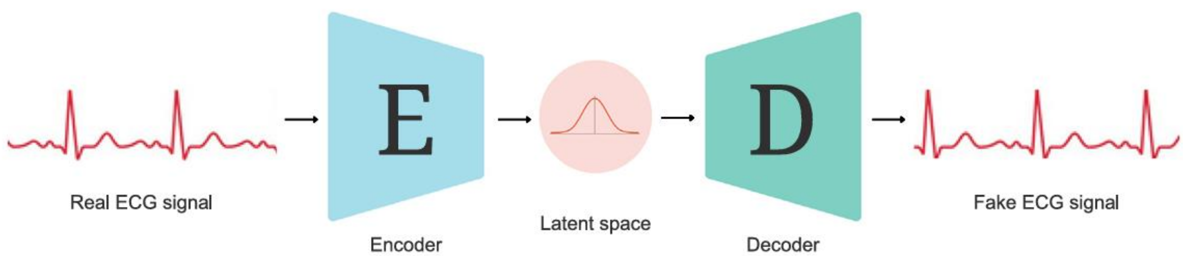


Figure 2*. A general design of VAE. In VAEs, the Encoder acquires latent variables and their posteriors from the original data. Next, values are extracted from the latent space and used as input for the Decoder, which then aims to recreate the original data (15)*

2.3.3 Diffusion Models (DMs)

Among the selected records, two referenced Diffusion Models: one discussed their application in generating image data, while the other described their use in synthesizing ECG signals. Research on synthesizing ECG signals with DMs remains limited, while the generation of images using this approach appears to be more prominent. In general, Diffusion Models are probabilistic generative models that combine two processes. In the first step, a Markov chain of diffusion steps is defined, where noise is gradually introduced to the input. Then, the model is trained to reverse this process, producing new data by starting from the noisy inputs. Figure 3 provides a visual summary of this introduction to Diffusion Models, highlighting their basics in generating ECG signals. In Figure 3, the two steps are characterized as the forward diffusion and the reverse diffusion.

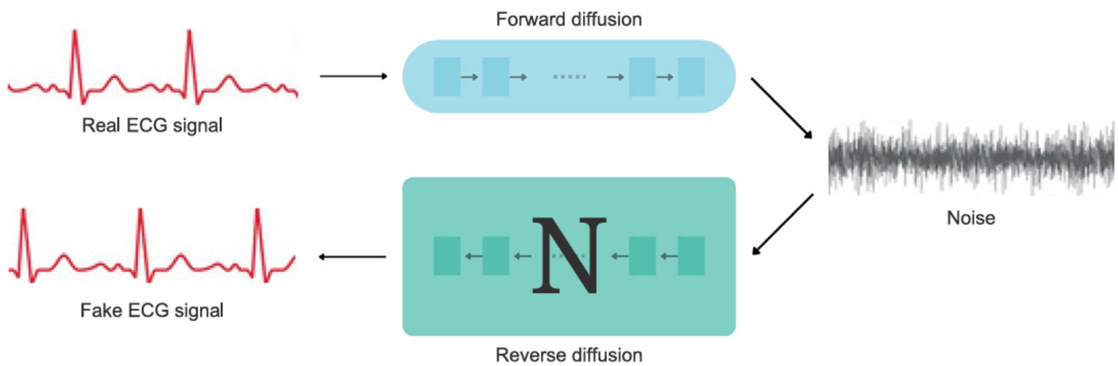


Figure 3*. A general design of DM. In the Forward diffusion, noise is added to the input. In the Reverse diffusion, SD is generated from the noisy input (15)*

2.3.4 MDClone Platform

Not utilizing deep learning like the three above-mentioned methods, MDClone uses statistical techniques to model real data directly without predefined distributions. In the article of Echo Wang et al., fictitious patient records were synthesized using the MDClone system with strong attention to ensuring privacy. The synthesizing process involves, firstly, data extraction, where real patient data is queried from the hospital's EMR data lake, integrating multiple data sources. Then, the synthetic data is generated using a multivariate algorithm that considers the relationships between variables. Again, it does not assume predefined distributions but instead derives them from the actual data. If certain subpopulations of the synthetic dataset are too small and could lead to patient identification, the system censors these values, further guaranteeing complete privacy. According to Echo Wang et al., deep learning models such as medGAN skip the longitudinal nature of medical records unlike MDClone, and compared to MDClone, medGAN does not emphasize privacy preservation to the same extent. (1)

3. METHODOLOGY AND MATERIALS

This literature review was carried out under the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (19). The following sections describe the search strategy, the selection of articles, and the search results.

3.1 Search and Select Process

Two databases served as the main literature resources for this study: Andor and PubMed. Andor provides access to Tampere University Library's printed and electronic books and journals, Tampere Universities' theses and dissertations, international articles, as well as databases, videos, and other publications. Students of Tampere University can access licensed materials remotely via their student accounts. (23) PubMed is a free online database of biomedical and life sciences literature, maintained by the U.S. National Library of Medicine (NLM) at the National Institutes of Health (NIH). It provides access to millions of research articles, including abstracts and full texts (when available), from journals covering medicine, biology, biotechnology, and related fields. (24) With these databases, the adopted search phrases were: "synthetic health data AND data quality assessment", "synthetic health data quality assessment", and "synthetic health data generation". The Boolean operator "AND" was used to narrow down the search by combining the terms suitably.

The search and select process was divided into three stages: Identification, Screening, and Inclusion. In the Identification stage, several filters were applied after entering the search phrases, for example, "Available online", "Open Access", "Free Full Text", "English", "Health Informatics", and "peer-reviewed". In the Screening stage, two levels of screening were applied: Title screening and Abstract screening. Importantly, in this stage, articles in which the main subjects were algorithms or approaches trained by synthetic data should be removed as they were not the focus of this research. Articles that were marked as "retracted" or repeatedly found were also excluded. The remaining articles, combined with articles found from citation searching, resulted in 54 articles included in the final review. Figure 4 illustrates the search and select process in the form of a PRISMA diagram.

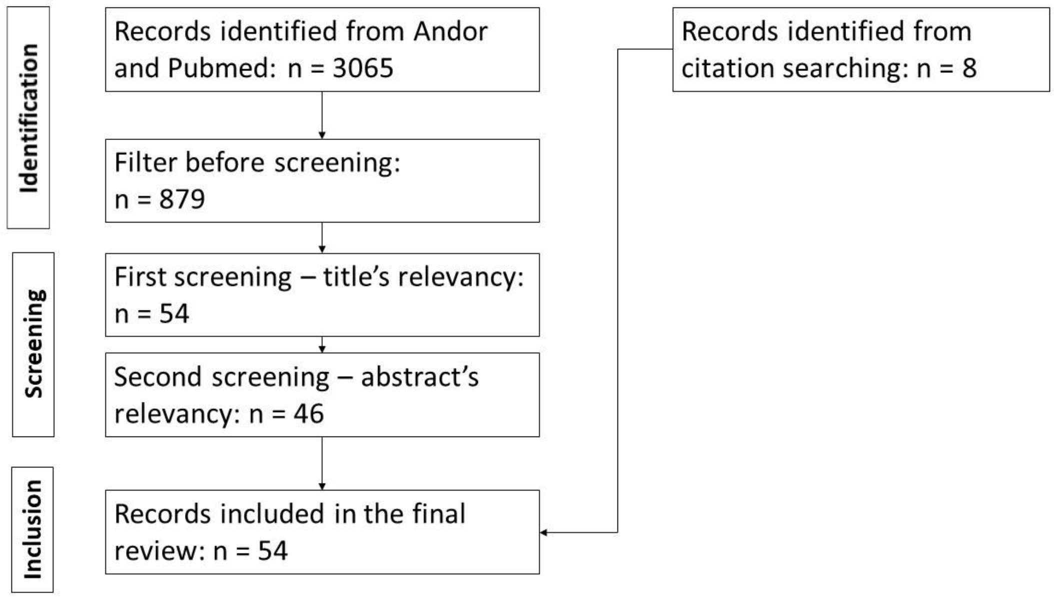


Figure 4. *PRISMA diagram*

3.2 Search Results Analysis

The selected articles were grouped based on their focus type of data: biomedical signal, medical image, and tabular EHRs. Besides that, articles that did not concentrate on one of these three types were placed in a separate group (namely "miscellaneous"), resulting in four groups of articles. Next, for each article group except the miscellaneous group, a mind map was created to organize the articles systematically, locating articles with high inclusiveness and those that were more specific. Notably, mind map drawing required careful and effective full-text screening of the articles to identify their objectives and topics correctly. The Zotero software was employed to store and arrange the selected articles according to the previously mentioned approach (25).

4. QUALITY DIMENSIONS OF SYNTHETIC HEALTH DATA

As mentioned in Chapter 2, “Theoretical Background”, evaluating the goodness of synthetic health data is vital in developing this state-of-the-art technology. Although different records use distinct approaches to assess synthetic data, common ground can still be found across the reviewed studies in the literature review process. Universally, synthetic health data can be evaluated by three main quality dimensions: Realism, Utility, and Privacy Risk. Several studies selected for this literature review present highly specific and insightful metrics and methodologies for evaluating synthetic data. However, upon completing the review, only the works of E. Budu et al. (17), Hernadez et al. (13), and G. Monachino et al. (15) were found to be sufficiently comprehensive and aligned with the scope of this thesis. While E. Budu et al. and Hernadez et al. covered synthetic EHRs and tabular data, G. Monachino et al. examined synthetic ECG signals. Consequently, the following subchapters, which delve into the quality dimensions, primarily draw upon these studies. Notably, the study of E. Budu et al. systematically categorized evaluation approaches based on data type (discrete, continuous, or type-agnostic) and application mode, making it a particularly well-structured and accessible reference. Therefore, it serves as the foundational groundwork for introducing each quality dimension in synthetic EHRs, with the work of Hernadez et al. providing supplementary details.

Besides that, Fairness and Carbon Footprint, explored by Vallevik and their team (10), are two additional quality dimensions discussed in this chapter. Although they are not going to be analyzed thoroughly compared to other quality dimensions, excluding them would result in a fairly incomplete consideration.

4.1 Realism

Realism, also called fidelity or resemblance in some research, refers to the ability of synthetic data to portray real data (13). Evaluating the realism of synthetic data reveals the degree of similarity between the samples in the synthetic data and the real data (17). To ensure completeness and clarity, the approaches applied to measure realism are grouped based on the nature of synthetic data.

4.1.1 Realism in Synthetic ECG Signals

Biomedical signals do not include only ECG signals, but the versatility of the following methods allows them to represent this class of synthetic health data. The first method is visual inspection, during which an expert cardiologist qualitatively assesses the morphology, rhythmicity, and other characteristics of the synthetic ECG signal. To reduce the subjectivity and make this approach less qualitative, the cardiologist would perform a real-vs-fake classification with a set of real and generated data, presented in random order and without giving feedback (by announcing the true answer during the classification). Evaluating based on quantitative metrics is another method in the study of G. Monachino's team. They divided the quantitative metrics into two groups. The first group includes metrics that evaluate the distribution of signal features (e.g., the amplitude) regardless of time order. Examples include Maximum Mean Difference (MMD), Relative Entropy (RE), Statistics of Features Distribution (SFD), and t-distributed Stochastic Neighbor Embedding (t-SNE) visualization. For instance, researchers can measure the amplitude, duration, or axis of specific ECG waves or fiducial points (e.g., RR interval, QT interval) to assess the plausibility of these values and their similarity to the corresponding distributions in the real data. The second group embraces metrics that compare pairs of signals, considering temporal order. Examples include Euclidean Distance (ED), Mean Absolute Error (MAE), Structural Similarity Index (SSI), Mean Square Error (MSE), Time Warping (TW), Frechet Distance (FD), and Chamfer Distance (CD). To put these metrics into practice, most studies compared each pair of real-fake signals. Additionally, to speed up the evaluation process, an automated real-vs-fake classifier is a proposed solution. However, the reliability of the evaluation heavily depends on the classifier adopted. (15)

4.1.2 Realism in Synthetic EHRs

E. Budu et al. compiled six methodologies: Multivariate methods, Comparison of Pairwise Correlations, Comparison of data statistics, Comparison of Variable Distributions, Similarity and Distance-based methods, and Clinical Validation Checks (17). Beyond that, Hernadez et al. came up with three analyses to evaluate synthetic tabular data's resemblance, which are Univariate Resemblance Analysis (URA), Multivariate Relationships Analysis (MRA), and Data Labeling Analysis (DLA). Uniquely, for each of these analyses, Hernadez's team gave a numerical score between 1 and 3 for the synthetic dataset, and after all three analyses were performed, a weighted average was computed to obtain a total resemblance. An example of weighting could be URA = 0.4, MRA = 0.4, and DLA = 0.2, and after applying these weights, the total score is calculated and rounded to the nearest integer. It is then classified as "Excellent" if the rounded value is 3, "Good" if it is 2, and "Poor" if it is 1. (13) In another study, Achterberg et al. divided the realism-evaluating methods into feature-wise similarity and similarity across features and the sequential (or other) dimension (5). While feature-wise similarity compares statistics like mean or variance of individual features in the SD to the corresponding ones in real data, the remaining type of similarity considers relationships across features or over time and checks if dependencies between features or patterns over time are preserved in SD (as in real data). The following paragraphs further explore some prominent metrics in the six methodologies utilized in E. Budu's research, which are summarized in Figure 5, inspired by E. Budu's work.

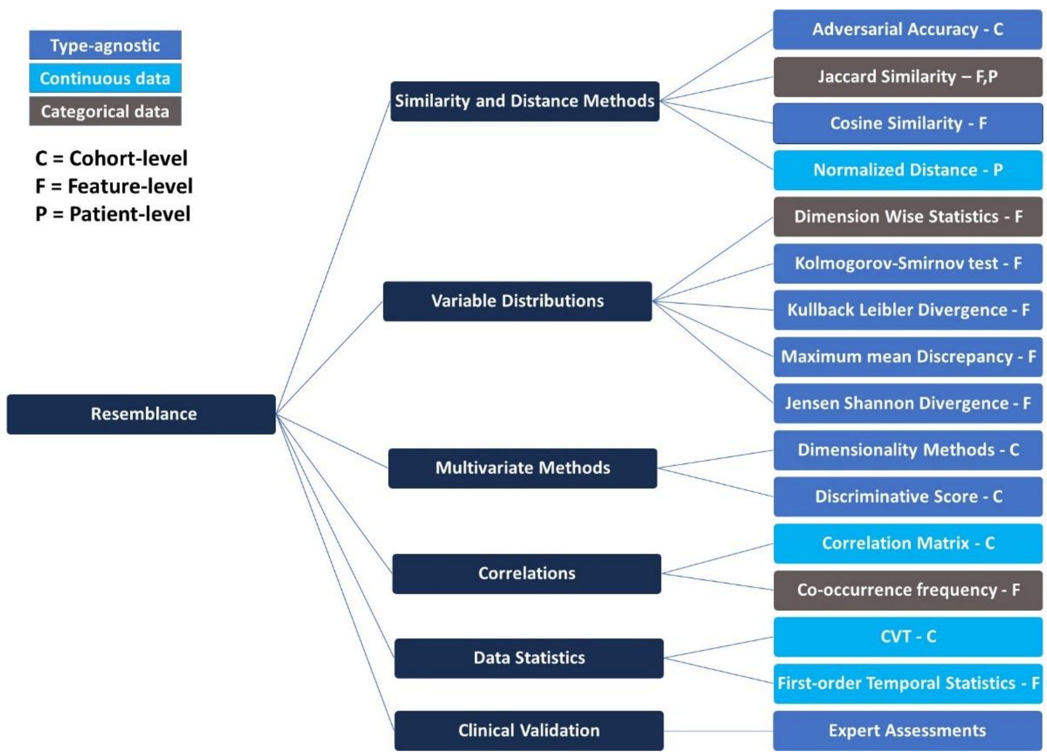


Figure 5. *Fidelity evaluation measures for synthetic EHRs are classified according to the data type: discrete, continuous, or type-agnostic. The measures are applied at different levels: across all patient records (cohort level), to individual variables (feature-based), or to specific instances of patient records (patient level) (17).*

Multivariate methods assess dataset fidelity by considering variable interdependencies, employing dimensionality reduction and discriminative scores. The methods described here can analyze data with continuous and discrete values and can be used to assess data at the cohort level. Dimensionality reduction methods transform high-dimensional data into a lower-dimensional space for visualization, allowing comparison between compressed and intuitive representations of synthetic and real data in the original feature space. (5,17) Dimensionality reduction methods can be named, such as, first of all, Principal Component Analysis (PCA), a linear technique that transforms data into principal components, preserving maximum variance. t-distributed Stochastic Neighbor Embedding (t-SNE) projects high-dimensional data into lower dimensions by creating probability distributions and minimizing the difference between the two distributions of the data in low and high-dimensional space. Uniform Manifold Approximation and Projection (UMAP) constructs a topological representation of data in a high-dimensional space and then projects that map into lower dimensions to minimize the distance between the data points and minimize the cost function. Non-metric Multi-Dimensional Scaling (NMDS), a variant of Multidimensional Scaling (MDS), aims to preserve the distances between points when mapping data to lower dimensions. Besides these dimensionality reduction methods, discriminative scores can be applied to quantify similarity by training a classifier to distinguish real and synthetic data, with higher classification errors indicating higher similarity. (17)

Correlation matrices are used to assess the correlation between continuous variables in both synthetic and real datasets. A matrix of correlations on the same variables is generated and compared for the synthetic and real EHRs. If all or most correlations in the real are captured in the synthetic data, then there is a similarity between the real and synthetic data. Alternatively, the pairwise correlation difference (PCD) can obtain a numerical correlation value. To quantify the difference in correlations, the L1 or L2 norm is applied to the difference between the correlation matrices, resulting in the PCD. A PCD close to zero signifies a high similarity between the real and synthetic data. Another method uses a co-occurrence frequency heatmap, visualizing how often different discrete variables appear together in both synthetic and real datasets. Similar patterns in the heatmaps suggest good synthetic data fidelity. (17)

Statistical fidelity measures compare summary statistics (e.g., mean, median, standard deviation) between real and synthetic electronic health records (EHRs). For continuous variables, methods like comparing means and standard deviations or using Constraint Violation Tests (CVT) on medians, minima, and maxima assess fidelity by checking if synthetic and real data statistics match. For a specific discrete variable, the mean and standard deviation can be computed for another continuous variable that is associated with this discrete variable. Then, the statistics between the real and synthetic datasets are compared to evaluate fidelity. A larger difference indicates lower fidelity. Additionally, first-order temporal statistics assess how well the synthetic data preserves correlations between time-dependent variables. (17)

Variable distributions examine whether variables in synthetic data and real EHRs come from the same distribution. This can be assessed by comparing histograms and distribution plots or using distribution similarity measures. E. Budu et al. proposed the Kullback–Leibler divergence (KLD), Kolmogorov–Smirnov (KS) test, Jensen Shannon divergence (JSD), Dimension-wise statistics (DWS), and Maximum mean discrepancy (MMD) to measure distribution similarity. (17) However, only KLD and JSD are discussed in detail due to their appearance in the research of Hernadez et al. (13). Apart from DWS, the rest of the distribution measures can be applied to both discrete and continuous data. Here, variables *P* and *Q* represent the distributions of a single variable in real and synthetic EHRs, respectively. KLD measures how much the probability distribution at the feature level of the real EHRs diverges from that of synthetic EHRs. The KLD score, D_{KL}, is calculated by integrating the product of each event's probability in P with the logarithm of the ratio between the probabilities in *P* and *Q*. Equation (1) illustrates the calculation of D_{KL} as follows:

$$D_{KL}\,(\mathrm{P}|\mathrm{Q}) = \sum_i P(i)\log(\frac{P(i)}{Q(i)})\;(17). \tag{1}$$

Here, *P(i)* and *Q(i)* are the probabilities of event *i* in the real and synthetic datasets. A KLD value of zero suggests that the distributions are similar, while higher values indicate more differences. Jensen-Shannon divergence (JSD) measures the similarity between two probability distributions. It works by averaging the KLD between the two distributions in both directions: from the real data distribution *P* to the synthetic data distribution *Q* and vice versa at the feature level. The JSD score, D_{JSD}, is between 0 and 1, where 0 means the two distributions are the same, and 1 means the two distributions are completely different. Equation (2) displays JSD:

$$D_{JSD}(P|Q) = \frac{1}{2}D_{KL}\left(P\middle|\frac{P+Q}{2}\right) + \frac{1}{2}D_{KL}(Q|\frac{P+Q}{2})\;(17). \tag{2}$$

Similarity and distance methods use a similarity metric or distance function to assess the accuracy of synthetic EHRs. E. Budu et al. introduced four measures: Cosine similarity (CS), Jaccard similarity (JS), Adversarial accuracy (AA), and Normalized distance (ND). Only CS and AA are reviewed in-depth because they are type-agnostic and can be used with both continuous and discrete data. Cosine similarity (CS) measures the angle between data points in real and synthetic EHRs. It compares vectors of values from both datasets. A score of 1 means high similarity, while −1 means dissimilarity. For synthetic data, a score between 0.4 and 0.8 is ideal, as it shows similarity without the datasets being identical. Equation (3) describes the CS, with *R* and *S* are vectors of values from variables in the real and synthetic EHRs, accordingly. R_i and S_i are the *i*th elements of the vectors *R* and *S*, respectively, and n is the total number of dimensions:

$$\cos(\theta) = \frac{R.S}{\|R\|\|S\|} = \frac{\sum_{i=1}^{n} R_i S_i}{\sqrt{\sum_{i=1}^{n} R_i^2}\sqrt{\sum_{i=1}^{n} S_i^2}} \text{ (17).} \quad (3)$$

Adversarial accuracy (AA) checks if synthetic data is similar to real data by seeing if a classifier can tell them apart. To compute AA, the first step is to draw empirical samples from the real and synthetic data distributions, *P* and *Q*, correspondingly. Equation (4) defines AA:

$$AA_{QP} = \frac{1}{2}(\frac{1}{n}\sum_{i=1}^{n} 1\left(d_{QP}(i) > d_{QQ}(i)\right) + \frac{1}{n}\sum_{i=1}^{n} 1\left(d_{PQ}(i) > d_{PP}(i)\right)), \quad (4)$$

where $d_{PQ}(i)$ is the distance between a point from the real distribution and its nearest neighbor in the synthetic distribution, $d_{QP}(i)$ is the distance between a point from the synthetic distribution and its nearest neighbor in the real distribution. Besides that, $d_{PP}(i)$ (for real data) and $d_{QQ}(i)$ (for synthetic data) are calculated as the distance from a sample *i* to its closest neighbor, excluding itself, within a group of *n-1* samples from the same distribution. The calculation counts how often the distance to the other group (real to synthetic or synthetic to real) is bigger than the distance to the closest point in the same group and then averages these counts. An AA value of 0.5 indicates “balanced accuracy” for the synthetic data, while AA > 0.5 means that the SD is noticeably different from the real data. (17)

Clinical Validation Checks is an evaluation approach that uses domain expertise from medical professionals, including clinicians, to verify the validity of synthetic EHRs. Clinical knowledge establishes rules that help eliminate inconsistent records from the synthetic data, ensuring the dataset retains its clinical relevance and plausibility (17). Besides synthetic EHRs, clinical validation checks can also be applied to synthetic biomedical signals and synthetic medical images (15,20).

4.2 Utility

Utility, also known as Usability, is the measurement of the usefulness of synthetic data. Evaluating the measurement of synthetic data helps answer the question: Can synthetic data be used for the same tasks as real data? (17) In the context of synthetic ECGs, G. Monachino et al. only explored the utility assessment for classification tasks. Particularly, the accuracy of a classifier can be used as an indirect measure to evaluate the quality of the generated data. In data augmentation applications, this involves assessing the improvement in classifier performance when synthetic ECG signals are added to the training set, providing a straightforward way to measure the effectiveness of the augmentation technique. Similarly, classifier performance is compared between anonymized and original datasets in data anonymization applications to determine if the essential ECG features required for the classification task are preserved. (15)

The following subsections elaborate on E. Budu's research on six approaches to assess the utility of synthetic EHRs, which are Predictive modeling, Dimension-wise predictions (DWP), Association Rule mining, Latent Space measures, First order proximity (FOP), and Cross-type conditional distribution (CCD). Figure 6 systematically categorizes these approaches as follows:

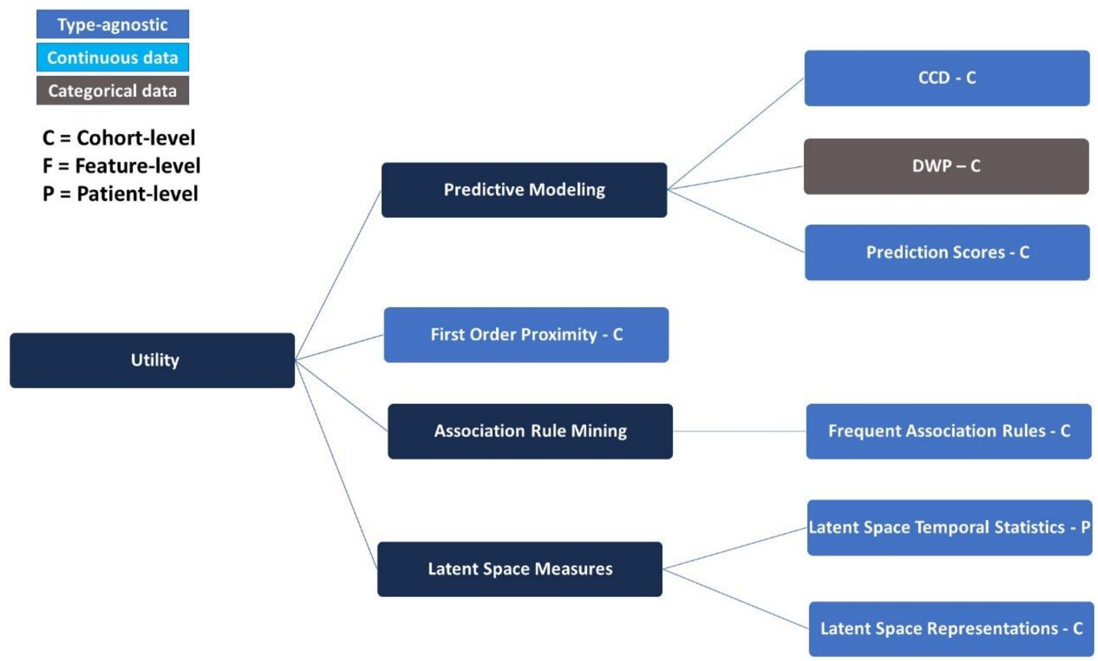


Figure 6. *Utility evaluation measures for synthetic EHRs are classified according to the data type: discrete, continuous, or type-agnostic. The measures are applied at different levels: across all patient records (cohort level), to individual variables (feature-based), or to specific instances of patient records (patient level).* (*17*)

4.2.1 Predictive Modeling

Predictive modeling evaluates the utility of synthetic EHRs by investigating ML models trained on both real and synthetic datasets. This approach utilizes metrics such as accuracy, the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), F1 score, and Mean Squared Error (MSE) to assess the applicability of synthetic data compared to real data. The choice of metric depends on the specific modeling task. It is worth noting that before testing the quality of synthetic data, the ML models must be reliable and built correctly. This can be ensured by first training and testing the models with real data (Train on Real, Test on Real (TRTR) framework). Researchers need to check that the performance metrics (e.g., accuracy, AUROC) are reasonable and consistent with expectations for the task. (17)

Typically, the evaluation follows the Train on Real, Test on Synthetic (TRTS), or Train on Synthetic, Test on Real (TSTR) framework, where models are trained on one kind of data and tested on the other (17). The TRTS framework checks if models trained on SD can make as accurate predictions as when the models are trained on real data, and to quantify and categorize the results of the evaluation, thresholds for the metrics differences between TRTS and TRTR are needed, as proposed by Hernadez et al. (13). They defined that utility is “Excellent” if the differences between metrics such as accuracy, precision, recall, and F1 score don’t exceed 0.2, “Good” if the differences are between 0.2 and 0.8, and “Poor” if the differences are higher than 0.8. On the other hand, the TSTR checks how well the SD captures the underlying patterns and relationships of the real data. In the experimental evaluations conducted by E. Budu et al., thresholds for metrics were not explicitly defined, but they indicated that an F1 score of 0.52 indicates that the “synthetic EHRs are unrealistic enough to train a classifier” (17).

4.2.2 Dimension-wise Predictions (DWP)

DWP assesses the extent to which a generative model captures the cross-dimensional relationships present in real data. This evaluation involves training two machine learning models, typically predictive models, where one is trained with real data and one is trained with synthetic data. There are identical numbers of data points from both real and synthetic datasets. The models are then tested on real data, so for each test instance, there are two prediction scores: one from the model trained on real data and one from the model trained on synthetic data. These prediction scores are compared using a scatter plot. In this plot, each test instance is represented by a point, with its x-coordinate representing the performance on real data and its y-coordinate representing the performance on synthetic data. If the points in the scatter plot align along a diagonal line, the synthetic data maintains a utility similar to the real data. This diagonal alignment represents an ideal scenario where the synthetic model's performance matches that of the real data model. (17)

4.2.3 Association Rule Mining

Association rule mining has also been used to evaluate the effectiveness of synthetic Electronic Health Records (EHRs). This approach examines whether the relationships between variables are maintained. For instance, association rules can be applied to determine if the relationships between pairs of variables are preserved between real and synthetic datasets. Frequent Association Rules (FAR) can be used for categorical variables to explore how associations between medical conditions at the record level in real EHRs are reflected in synthetic EHRs. The utility is measured by comparing the association rules ratio found in synthetic and real EHRs. (17)

4.2.4 Latent Space measures

According to E. Budu et al. (17), Yan et al. and Zhang et al. used Latent Space Representations (LSRs) to evaluate whether synthetic data preserves the structural properties of real data. The approach compares the distributions of variances in the latent space between the real and synthetic datasets. Specifically, a β-Variational Autoencoder (β-VAE) is trained on the real data to obtain its latent space representations and corresponding variance distribution. Then, the SD is input into the trained β-VAE to obtain its variance distribution on the latent dimensions. The distance between the mean of these variance distributions is then measured for each latent dimension. Besides that, if the variance distribution of the synthetic data is closer to 1, a loss of important structural properties has been observed, leading to reduced utility. Another method involves comparing the distributions of latent variables obtained through Singular Value Decomposition (SVD) on both datasets. Matching distributions indicate that the synthetic data maintains the structural properties of the real data. (17)

4.2.5 First-order proximity (FOP)

FOP is used to check if synthetic data keeps the local structure of real data, especially for discrete variables like medical codes. In this context, local structure means the way certain variables are related or tend to occur together in the data. FOP works by, firstly, creating a graph where nodes are variables and edges show the co-occurrence frequency of variables in the population. The FOP is the difference between distances in the graphs of equal portions of the synthetic data, *s*, and real data, *r*, when a function *h(.)* is employed to normalize the distance values. Equation (5) defines the FOP as follows:

$$\sum h\left(\frac{|FOP(s)-FOP(r)|}{FOP(s)+FOP(r)}\right). \tag{5}$$

A small difference, either minimal or falling within the range of the distances in a real vs real setting, means the graphs are very similar, so the synthetic data preserves the local structure of the real data well. (17)

4.2.6 Cross-type conditional distribution (CCD)

CCD assesses whether a generative model can maintain the conditional distribution of features in the real dataset. This can be calculated by computing the distribution of one feature on another. Additionally, this approach can be done by comparing the mean and standard deviation of each conditional distribution between real and synthetic EHRs. For example, in the work by Yan et al., "Generating Electronic Health Records with Multiple Data Types and Constraints", the demographic distributions for each health condition and procedure were compared. (17)

4.3 Privacy

Privacy assessment tells whether synthetic data leaks information from real data or not by evaluating the extent to which the generated data differs from the training samples. Assessing the privacy of SD helps answer the question: Given a sample of synthetic data *S* and a sample of real data *R*, is it possible to identify which information in *R* was used to generate records in *S*? (17) As already mentioned in Chapter 2, "Theoretical Background", privacy preservation is one of the main reasons for the emergence of synthetic health data, so, privacy is evidently an important check in assessing SD's quality. It examines whether SD can be utilized in real-life studies while ensuring compliance with privacy laws. (17)

In ECG data anonymization, G. Monachino et al. addressed the utmost limitation of available works that proposed evaluation metrics for privacy purposes, with only three studies doing so. The first metric mentioned in Monachino's article is the membership inference attack (MIA), which Monachino's team could not evaluate further than its definition of how this metric could evaluate synthetic ECG. Hence, MIA is explained further in the following subsections about metrics to assess synthetic EHRs/tabular data. The second metric introduced by Monachino's team is the accuracy of the so-called "Verificator", in which a value close to 50%, i.e., random guessing, implies that the generated signal is anonymized. (15)

The subsequent subsections explore three privacy-related metrics of E. Budu et al., which are disclosure metrics, distance metrics, and the effect of differential privacy (DP) (17). Figure 7 categorizes these metrics as follows:

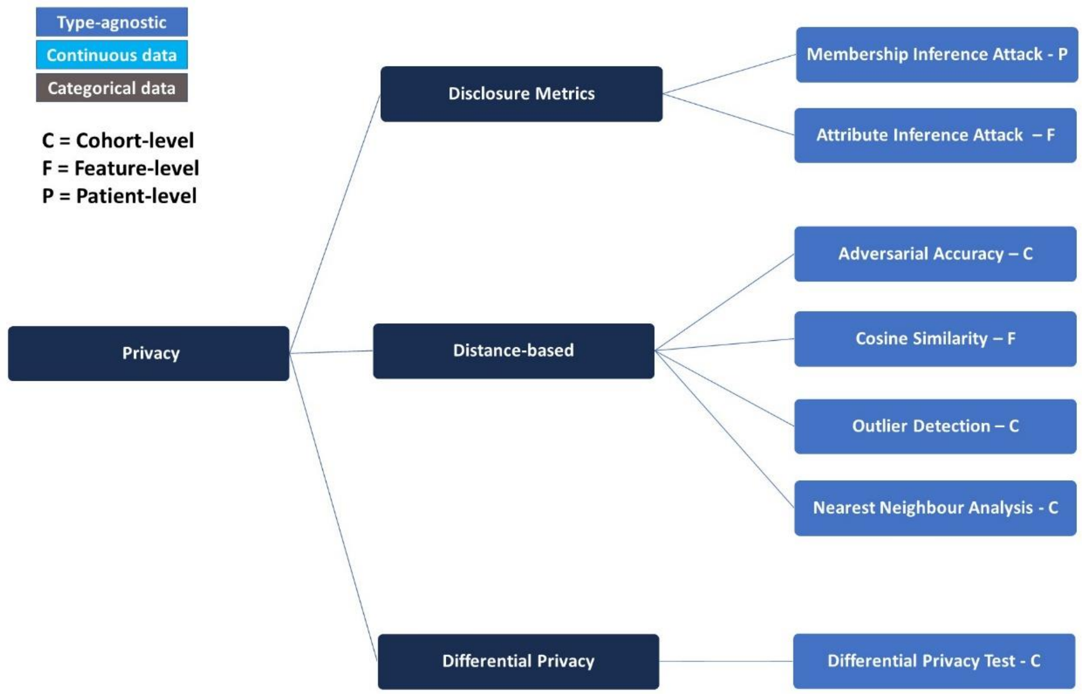


***Figure 7**. Privacy evaluation measures for synthetic EHRs are classified according to the data type: discrete, continuous, or type-agnostic. The measures are applied at different levels: across all patient records (cohort level), to individual variables (feature-based), or to specific instances of patient records (patient level).* (17)

4.3.1 Disclosure metrics

Disclosure metrics are divided into two types: unique identity disclosure and sensitive attribute disclosure. Unique identity disclosure ensures that an individual's identity cannot be identified from the synthetic data, whereas sensitive attribute disclosure prevents specific personal information from being revealed through the synthetic data. To evaluate these disclosures, two tests are applied: a test on membership inference attack (MIA) and a test on attribute inference attack (AIA). While in MIA, the attacker aims to determine whether real patient records have been used to generate synthetic data. AIA indicates that the attacker has access to synthetic EHRs and a subset of features of some real data and is trying to guess the value of sensitive features. (13,17)

Hernadez et al. nominated in their study the process of how a hypothetical attacker performs an MIA. According to them, the attacker must be able to access the whole synthetic data set and a subset of real data (RD) randomly distributed. For each row in the RD subset, the attacker aims to recognize the closest records in the SD set. The attacker will conclude that there is at least one row from the SD set close enough to the RD if there is any distance lower than a certain threshold. Hernadez et al. simulated the attacker's investigation by suggesting the calculation of the Hamming distance, which is described by equation (6), between real and synthetic records:

$$haus\_dist(S, R) = max\{h(S, R), h(R, S)\}. \tag{6}$$

The Hamming distance serves as the proportion of non-equal attributes between synthetic and real records, so a lower value indicates more similarity between two records. Multiple thresholds, which are 0.4, 0.3, 0.2, and 0.1, are applied to assess similarity. The attack's success rate is evaluated by computing the accuracy, which is the proportion of correct predictions made by the attacker, and precision, the proportion of records used for generating the SD identified by the attacker, for each threshold. Privacy is categorized as "Excellent" if accuracy and precision are ≤ 0.5 for all thresholds, indicating low re-identification risk. If they exceed 0.5 for some or most thresholds, privacy is rated as "Good" or "Poor," respectively, reflecting increased disclosure risk. (13)

In an AIA, Hernadez et al. suggested that the attacker use ML models trained with SD to predict the values of unknown attributes of the RD records. In this process, the attacker needs to, first, define the quasi-identifiers (QIDs), which are features included in the known features of the real data, of each data set. During the training of the ML models, QIDs are used as input features, and non-QID attributes are the target values. The models intend to learn the pattern of the SD set and predict the anonymous features of the real patients. Data batches are generated with each QID combination, and each batch is evaluated using accuracy for categorical attributes and root-mean-square error (RMSE) for numerical attributes. Privacy risks are exposed if accuracy is high while RMSE is low. If the attacker correctly predicts more than 60% of the hidden attributes, the privacy risk is rated as "Poor." If the percentage of the correctly predicted attributes ranges from 60% to 40%, the privacy risk is "Good", and a percentage lower than 40% indicates "Excellent" privacy preservation against AIA. (13)

4.3.2 Distance metrics

Distance metrics, which are already brought up in realism validation, can also be utilized to measure privacy risk. However, contrary to realism validation, the goal of the privacy test is that the synthetic EHRs must not be too similar to the real EHRs. Cosine distance, which is computed as 1 minus cosine similarity, is an approach representing distance metrics. A cosine distance of 0 indicates similarity, and thus, privacy is not preserved, while the opposite holds. Adversarial accuracy is also an approach for distance metrics, which is already described in the realism check. To employ AA in privacy check, the privacy score is computed by first taking the AA score between the synthetic and real training data, AA_{Train}, and the AA score between the synthetic and real test data, AA_{Test}. Then, the privacy score equals AA_{Test} subtracted by AA_{Train}, and privacy is preserved if the privacy score is minimal. The third approach for distance metrics found by E. Budu et al. is Outlier Detection with Density-Based Spatial Clustering of Applications with Noise (DBSCAN). This approach checks if the real and synthetic datasets have the same outliers. If there are no matching outliers, privacy is protected. (17)

4.3.3 Effect of Differential Privacy (DP)

Differential privacy is a mathematical framework that protects individual data privacy by introducing controlled noise to the data. This concept ensures that the probability of a specific outcome (a result produced by analyzing a dataset) is minimally affected by the presence or absence of a particular record in the dataset. Equation (7) describes DP where *D* is the dataset being analyzed, *D'* is the dataset where one record has been added or removed, *P* stands for probability, ϵ and δ define the privacy bounds (smaller values indicating stronger privacy) *M* is a randomized algorithm, and *S* is a subset of possible outcomes when *M* is applied to *D* or *D'*:

$$P[M(D) \in S] \leq e^{\epsilon} \times P[M(D') \in S] + \delta. \tag{7}$$

Equation (7) says that the probability of getting a certain result when using dataset *D* should not change too much if *D* is replaced with *D'* (where one record is added/removed from D). DP is not a direct approach to measuring the privacy of generated data, but it is used when creating synthetic data to minimize privacy risks. To measure privacy preservation, SD users can measure the realism and utility of the SD generated with DP and compare them with the results when DP is not included. Reporting different privacy bounds used when assessing the realism or utility of SD is another way to tell whether privacy is protected. (17)

4.4 Fairness and Carbon Footprint

The omission of Fairness and Carbon Footprint, interesting quality dimensions discovered in Vallevik's study (10), would potentially make the results of this chapter rather insufficient. Moreover, having this subchapter implies that these two quality dimensions deserve the same level of attention as the previously mentioned ones. Indeed, Vallevik et al. indicated that the number of studies on these two dimensions is extremely limited compared to those on the abovementioned dimensions. Thus, a detailed analysis of Fairness and Carbon Footprint within this subchapter is precluded; instead, a concise introductory overview will be provided.

Fairness ensures that synthetic data does not introduce or amplify bias, meaning that minority groups should achieve the same prediction performance as others. It also emphasizes transparency in the data generation process. Carbon Footprint measures the environmental impact of generating synthetic data, including the computational resources used for preprocessing, training, and inference. This can be assessed through factors like model size, computation time, energy sources, and storage methods. (10)

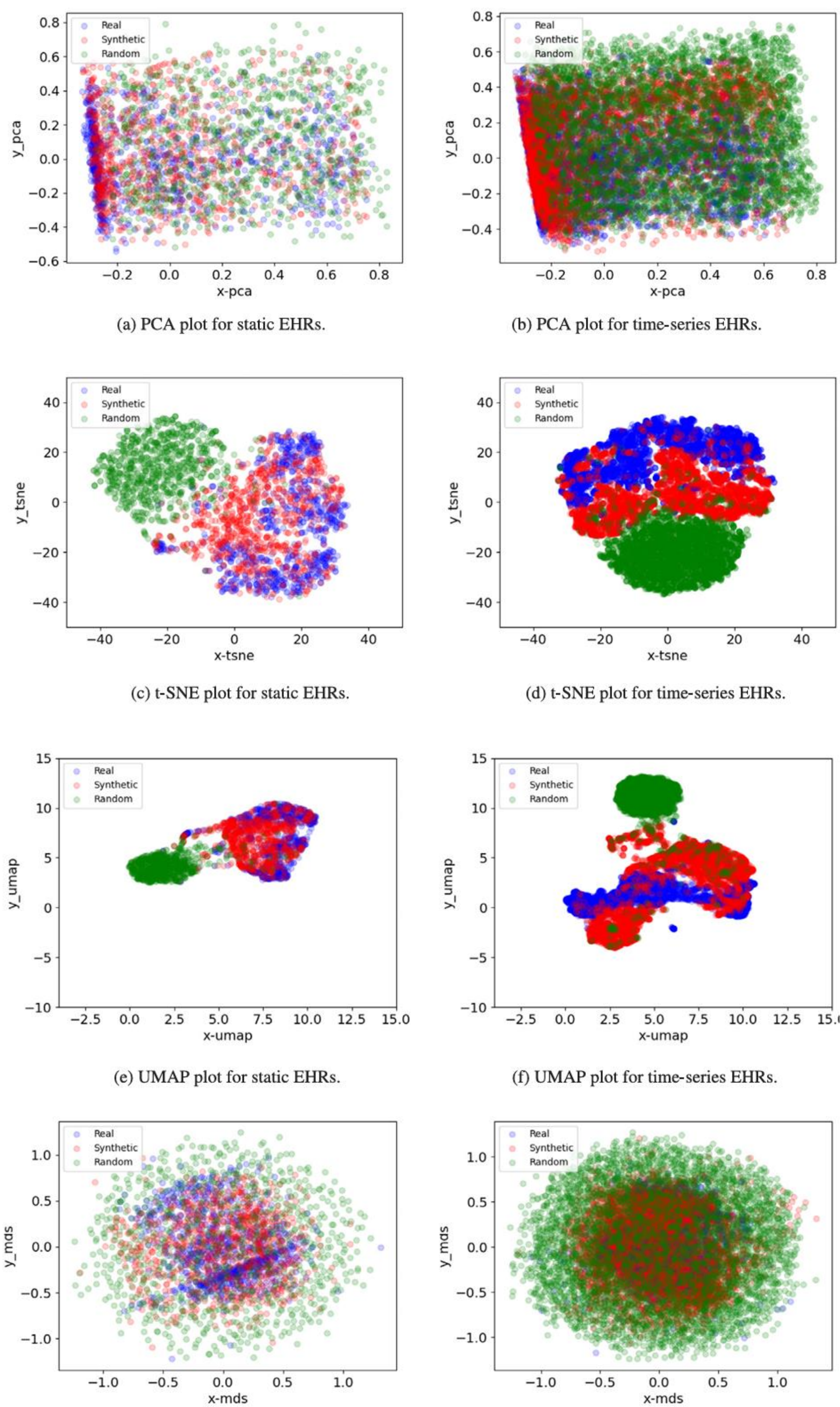
5. PRACTICAL QUALITY ASSESSMENT OF SYNTHETIC HEALTH DATA

While Chapter 4 offers metrics and methods to assess SHD's quality on theoretical ground, this chapter aims to provide practical evaluations that employ the presented quality dimensions. Since Chapter 4 is mainly built upon Budu's work, it is logical to include the experimental evaluations conducted by E. Budu in this chapter. Presenting the results of this experiment facilitates readers in deeply understanding the evaluation metrics and methods. This chapter concludes with a quality-evaluation framework, which considers the complete process of synthetic data generation and utilization, proposed by Vallevik's research team.

In Budu's experimental evaluations, the MIMIC-IV dataset was used to generate synthetic EHRs. MIMIC-IV (Medical Information Mart for Intensive Care-IV) is a de-identified database of over 50,000 adult patient admissions (2001-2012) from Beth Israel Deaconess Medical Center's critical care units. Budu's experiment employed a cohort of heart failure patients with variables such as age, length of stay, time from the previous admission, and sodium levels in the blood, to name a few. In this experiment, CTGAN was used to generate static EHRs, while TimeGAN generated time-series EHRs. (17)

5.1 Realism Evaluation

Budu's realism evaluations employed dimensionality reduction methods, which are PCA, t-SNE, UMAP, and MDS, on real and synthetic data in static and time-series forms, as in Figure 8. The visualizations in Figure 8 display real (blue), synthetic (red), and random (green) data in a lower-dimensional space. Random data is added to assess if the synthetic data matches real distributions or resembles noise. (17)



(a) PCA plot for static EHRs.

(b) PCA plot for time-series EHRs.

(c) t-SNE plot for static EHRs.

(d) t-SNE plot for time-series EHRs.

(e) UMAP plot for static EHRs.

(f) UMAP plot for time-series EHRs.

(g) MDS plot for static EHRs.

(h) MDS plot for time-series EHRs.

***Figure 8**. PCA, t-SNE, UMAP, and MDS visualizations of real and synthetic EHRs.* (17)

In each plot, the similarity between real and synthetic data is implied by the overlap of the distribution of data points from each set. Thus, Budu concluded that similarity was observed between static real and synthetic data in their experiment. However, on the time-series side, UMAP and t-SNE plots show insignificant similarities between the distribution of the data points. Budu emphasized that data from different patient visits were combined in the dimensional reduction in the time-series case, so the resemblance evaluation regarding temporal patterns was impractical. (17)

5.2 Utility Evaluation

Regarding the utility evaluation of synthetic data, E. Budu et al. employed the Prediction Scores and CCD metrics. First of all, the review of the Prediction Scores measurement is explained. Budu's experiment acquired the prediction scores from predictive tasks on the static and time-series EHRs. In the static case, they trained a Random Forest classifier to predict whether a patient would be readmitted to the hospital using the TSTR framework, and the classifier's performance was measured with the F1 score. For time-series data, they used the MAE score from a previous study, which trained a sequence prediction model (RNN) on artificial data to predict future sequences based on past sequences. This model was subsequently tested on real data to predict the next time step. E. Budu et al. reported that the F1 score for the static case was 0.52, meaning that the classifier's performance was only as good as random guessing. The higher the F1 score, the greater the utility of the synthetic data. In contrast, the time-series synthetic data was realistic enough to train a model that accurately predicts future values, with an MAE of 0.33 (lower MAE is better). This indicated that the synthetic data successfully reflects the patterns and connections in the original data, making it useful for building models. (17)

Budu et al. employed the CCD metric by determining the conditional distribution of the readmission flag on CCI (Charlson Comorbidity Index), a mortality indicator for patients. Their result illustrated that in real EHRs, patients had the highest probability of being readmitted (flagged as 1) within a CCI range of 2 to 4. A similar pattern appeared in the synthetic data, but the probability of readmission differed significantly from that of real data. This discrepancy in the distribution of patients with flags 0 and 1 suggested that the synthetic EHRs did not accurately reflect the conditional distribution of one variable given another, making them less reliable than real data for this purpose. (17)

5.3 Privacy Evaluation

On the Privacy aspect, Budu's research studied the use of the Outlier Detection method. The result, shown in Figure 9, was depicted by a 2D scatter plot generated through DBSCAN clustering, showing real and synthetic EHRs derived from CTGAN. The real and generated data were projected into a 2D space using PCA. This plot visualized the distribution of data points, distinguishing between those classified as normal and outliers by DBSCAN. (17)

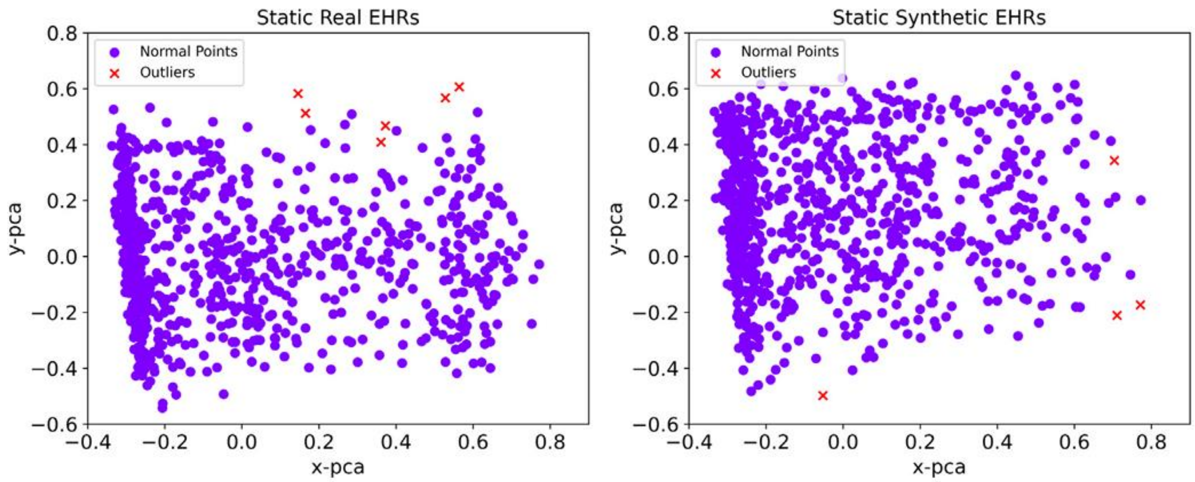


***Figure 9**. Outlier Detection using DBSCAN (eps = 0.1, $min\_samples$ = 5) was performed on real and synthetic static EHRs after dimensionality reduction with PCA. Outliers in real EHRs plot were distributed differently from outliers in synthetic EHRs plot, suggesting that privacy was preserved in synthetic data.* (17)

Via the plot, Budu concluded that the synthetic EHRs maintain the privacy-preserving ability due to the difference in outliers between the real and synthetic cases. However, Budu acknowledged that further in-depth evaluations were needed to accurately assess whether patient privacy in the real data was guaranteed. (17)

5.4 A Proposed Quality Evaluation Framework

Vallevik et al. did not provide any detailed studies on how the Fairness and Carbon Footprint of synthetic data were evaluated in real-world settings. However, they were able to propose a comprehensive framework that offered a structured, stepwise approach to assessing synthetic data quality. Their framework took into account the entire pipeline of SD generation and use, and it aligned with regulatory guidelines like the AI Act, the EU Medical Device

Regulation (MDR), and the U.S. Food and Drug Administration recommendations for AI/ML applications. In this framework, five quality dimensions were used: Usability (Utility), Similarity (Realism), Privacy, Fairness, and Carbon Footprint. These quality dimensions were not only applied to the synthetic data but also to the training data and the SD generator. Their framework was concisely analyzed according to the following six stages: (10)

- Priorities and Context (Stage 0): In this stage, the goal is to determine the intended use of synthetic data and set priorities for quality dimensions. SD examiners can consider any variables in the original dataset that are supposed to be more important than others for the expected use. (10)
- Training Data Quality Evaluation (Stage 1): The validity of SD inferences relies on the training data, thus requiring an initial check for representativeness, anonymity, bias, and preprocessing needs. Representativeness ensures relevance to the target cohort, while selection bias should be assessed to address minority underrepresentation. Record completeness ensures the dataset is suitable for use, resulting in preprocessing needed for missing data and outliers. (10)
- Generation Performance Evaluation (Stage 2): SD generation faces challenges such as algorithm convergence, mode collapse, mode invention, noisy data, and copying of training data, among others. This stage aims to assess the generative model's stability, bias handling, and overall quality. Generators can be tuned for specific data types (image, text, or structured data) or to address the previously mentioned issues. (10)
- Synthetic Data Evaluation (Stage 3): This stage evaluates the quality of synthetic data based on four dimensions: Similarity (Realism), Usability (Utility), Privacy, and Fairness. The methods and metrics from Budu's study presented in Chapter 4 can be utilized at this stage. Note that in Vallevik's study, Clinical Check is used at this stage to assess Usability, unlike in Budu's study where Clinical Check assesses Realism. (10,17)
- Analytical Validation (Stage 4): This stage focuses on the downstream performance of synthetic data. It emphasizes the Utility dimension of synthetic data, particularly on the Predictive Modelling task presented in section 4.2.1. (10)
- Clinical Validation and Monitoring (Stages 5 & 6): These stages target to confirm the real-world applicability of synthetic data, particularly the performance of the SD-trained AI tool in a real-world scenario. After implementation, performance monitor-

ing (Stage 6) contributes to Post-Market Surveillance. (10) Note that Clinical Validation in Stage 5 is placed in a downstream context, unlike Clinical Check in Stages 1 and 3.

Table 1, a simplified version of Vallevik's work, visually recaps this framework.

Table 1. *The quality assurance framework goes through the entire synthetic data pipeline, emphasizing key evaluations at each step*. (10)

Stage 0: Priorities and Context	**Stage 1: Training data evaluation**	**Stage 2: Generation process evaluation**	**Stage 3: Synthetic data evaluation**	**Stage 4: Analytical validation**	**Stage 5: Clinical validation**	**Stage 6: Monitoring**
Usability	Clinical Check: Representativeness	Generator performance	Clinical Check	Downstream performance	Downstream performance	Downstream performance
Similarity		Multivariate similarity (coverage)	Univariate similarity, Multivariate similarity			
Privacy	Privacy level	Privacy budget	Privacy level, Re-identification attack risk	Privacy budget		
Fairness	Bias	Bias	Bias	Bias and fairness, Transparency	Bias and fairness	Bias and fairness
Carbon Footprint	Manual processing needs	Manual processing, Computational footprint				

In Table 1, Stages 1 to 3 are the stages when synthetic data is produced, while Stages 4 to 6 are the stages when synthetic data is used in downstream applications. It is worth mentioning that in this framework, each stage has its exclusive considerations. To illustrate, when Usability (Utility) is evaluated in Stage 3, Clinical Check is considered, but in Stage 4, the Utility of synthetic data is evaluated by its downstream performance, for example, in Predictive Modelling. Another instance is the Privacy consideration in Stages 3 and 4. While in Stage 3, one must consider if the synthetic data points are too similar to the original or not, or the risk of an attacker inferring sensitive personal data from SD, Stage 4 considers the privacy risk when sensitive information is revealed in the downstream application.

6. LIMITATIONS IN SYNTHETIC DATA QUALITY ASSESSMENT

In Chapter 4, methods and metrics are given for the evaluation of three quality dimensions of synthetic data: Realism, Utility, and Privacy. Since Chapter 4 is primarily based on Budu's work (17), it makes sense that this chapter contains Budu's findings on the limitations of the methods and metrics. These limitations establish a crucial foundation for future research in synthetic health data. They address critical problems such as subjective judgment, the inherent dependencies between a patient's variables characteristic of synthetic data, the disregard for its time-series nature, and the shortcoming in testing its downstream applications. Tackling these issues can contribute to a more thorough and refined assessment of synthetic data.

6.1 Limitations in Dimensionality Reduction Methods

Although Dimensionality Reduction Methods allow comparison between compressed and intuitive representations of synthetic and real data, they do not calculate a numerical metric to compare the data, as seen in Figure 8. Interpreting low-dimensional visualizations can be challenging, making it difficult to determine how closely synthetic EHRs resemble real EHRs, as the assessment relies on subjective judgment. Budu et al. (17) emphasized the importance of incorporating a quantitative score derived from low-dimensional representations. Such a measure would help evaluate how closely synthetic EHRs resemble real data in a given representation, minimizing reliance on subjective judgment.

6.2 Limitations in Variable Distributions Measures

In real-world EHRs, various variables are often interdependent. For instance, a patient's diagnosis influences the prescribed medication or the lab tests they receive. However, most methods for analyzing variable distributions operate in a univariate manner, evaluating the quality of each variable independently. These approaches fail to emphasize the crucial dependencies between variables, making it difficult to determine whether the relationships present in real data are accurately reflected in the synthetic EHRs. E. Budu et al. created pairwise distribution plots in their experimental evaluations, which were not presented in Chapter 5, to assess the relationship between two variables, but they were limited to just two dimensions. (17)

Additionally, E. Budu acknowledged that their research lacked a thorough and systematic application of methods like distribution similarity measures. The reviewed studies used in Budu's research did not provide information on how distributions were represented before applying these similarity measures. Particularly, Budu et al. discovered that, in their experiment, changing the number of bins on the distribution of patient ages influenced the KLD and JSD scores obtained. Thus, it is crucial to report how the distributions are represented before applying similarity measures. (17)

6.3 Limitations in Distance-based Realism Methods

Distance-based Realism methods evaluate the similarity between feature vectors in real and synthetic datasets. Similar to variable distributions analysis, these methods operate at a univariate level, examining variables separately. However, relying solely on a univariate assessment is not a practical approach for determining the overall quality of a synthetic dataset. (17)

6.4 Limitations in Evaluating Time-Series Data

Budu et al. recognized that most existing methods disregard the temporal dependencies between patient encounters, leading to the lack of evaluation methods for time-series EHRs. In Budu's work, only the Statistics measures of the Realism test mention the validation of temporal statistics of synthetic data. As stated in Section 5.1, Realism Evaluation, when evaluating time-series data using Dimensionality Reduction Methods, different patient visits were combined, making the validation of temporal characteristics of synthetic data impossible. A more effective approach would be to assess multiple encounters simultaneously. However, this becomes impractical when encounters are large, as is often the case in real-world EHRs. Budu et al. suggested that evaluation metrics such as trajectory similarity measures, which assess time dependencies, should be used when evaluating temporal EHRs. (17)

6.5 Lack of Clinical Validation and Impact of Dataset Format

The displayed methods and metrics that evaluate SD's quality dimensions ignore the value of clinical validation. Evaluating the quality of synthetic EHRs should go beyond analyzing their structural and statistical properties; thus, it is essential to include assessments that verify the clinical relevance of the generated data. Incorporating the expertise of domain specialists into both the generation and evaluation processes is key to ensuring the creation of clinically meaningful synthetic EHRs, enhancing their utility by aligning them with real-world patient scenarios. (17) It is notable that Clinical Validation Check, a Realism method mentioned in Section 4.1.2, only inspects whether the synthetic EHRs look realistic or not. The lack of clinical validation in this section means the inspection of whether the synthetic EHRs are successfully applied in downstream clinical applications or not.

When selecting evaluation measures for synthetic EHRs, it is crucial to consider the dataset's format. EHRs can be structured or unstructured textual data, and standards like HL7-FHIR (Health Level 7, Fast Healthcare Interoperability Resources) provide guidelines for organizing structured EHR datasets, introducing specific characteristics to the data. When generating synthetic data that follows these standards, evaluation measures must reflect the characteristics imposed by the format. Consequently, the evaluation criteria for datasets structured according to HL7-FHIR may differ from those used for datasets with other formatting standards or no standards at all. Therefore, tailoring evaluation metrics to the dataset's format, an aspect neglected by the given methods and metrics in Chapter 4, is essential for a thorough assessment of synthetic EHRs. (17)

7. DISCUSSION

This thesis initially aimed to discover methods and metrics employed to evaluate the quality of time-series synthetic health data, such as ECG and EEG (electroencephalograph) signals. However, this objective was asserted to be impractical during the article search and select process due to the severe lack of available results. Moreover, the relevant studies related to this initial goal primarily evaluated each type of signal separately rather than assessing time-series data as a whole. Thus, the focus was expanded from time-series synthetic health data to encompass all types of synthetic health data, providing a broader perspective on the topic. As a result, the selected articles were divided into four groups: medical image data, biomedical signals, EHRs and tabular data, and the miscellaneous group. From that, attention shifted to the biomedical signals and EHRs and tabular data groups, as both can be reasonably classified as time-series data, whereas synthetic medical images were already extensively studied. Although the presented evaluation techniques for biomedical signals were rather limited, this thesis successfully demonstrated a relatively comprehensive and systematic approach to assessing EHRs and tabular data. The following paragraphs present some remarks about the results of this thesis.

One significant observation that emerged throughout the research process is the ambiguity in distinguishing between the concepts of "Utility" and "Realism." None of the selected studies explicitly addressed this issue, including the most recent ones that reviewed earlier research. Intuitively, “Utility” and “Realism” can be combined into a single quality dimension, as high utility in synthetic data inherently implies a degree of similarity to real data. This reasoning follows a logical progression, suggesting that synthetic data with greater utility should exhibit characteristics that closely align with those of real-world data. This opinion is somewhat supported by the study conducted by Murtaza and their partners, in which SD was assessed based on two aspects only: Realism, which was further divided into Resemblance and Utility, and Privacy (1). This alternative taxonomy could be potentially explored more, but unfortunately, the methods and metrics presented by Murtaza et al. are not as exhaustive as those employed in Chapters 4, 5, and 6. Additionally, in Chapter 4, certain methods for evaluating the utility of synthetic data, such as association rule mining and FOP, compare feature relationships in synthetic data to those in real data. At first glance, this approach appears similar to how the correlations method assesses the realism of synthetic data, suggesting a strong link between evaluating utility and realism. However, a distinction can still be made between Utility and Realism tests. Specifically, in the Realism test, Budu et al. frequently used the word family “visualize” to highlight the nature of these methods, as they primarily focused on assessing the visual representation of the data. On the other hand, Utility metrics assess how well synthetic data preserves the statistical properties and relationships necessary for analysis and modeling. They measure whether a model trained on synthetic data would perform similarly to one trained on real data. These metrics evaluate whether synthetic data can be effectively used for downstream tasks. (17)

Another important remark is the article search and selection process and the search results analysis. Although the research question of this study was defined at the beginning stage, no specific breakdown questions were explicitly determined to guide the search and select process smoothly and without hassle. As a result, determining whether an article was suitable for the final review often proved challenging. If a predefined set of 3-4 breakdown questions had been established, the selection process would have been smoother, more consistent, and better organized. Without these guiding criteria, the chosen articles covered an overly broad scope. Due to the scarcity of available research on this domain, articles were selected with a preference for including slightly irrelevant ones rather than risking the omission of potentially valuable studies. The hope was that even minor details in seemingly unrelated articles might provide useful insights. However, this approach led to difficulties in analyzing and extracting information from the selected articles. Besides these errors, it is worth noting that the search phrases were quite long, yet they were able to obtain more relevant results due to the scarcity of research in this domain. Additionally, the automatic citation and tag function of Zotero were extremely helpful in the article-analyzing and thesis-writing process.

Chapter 4 was built primarily on the study of Budu et al. due to the insightfulness of their metrics and methods. However, in their work, the border between high-quality SD and low-quality SD was not clearly defined systematically and consistently across the quality dimensions. Moreover, Budu et al. did not aggregate the results of individual measures to derive an overall score for each quality dimension. As a result, their approach made the methods and metrics less meaningful in practical applications. In contrast, Hernadez’s scoring system offers a potential improvement (13). In Hernadez’s study, each measure within a quality dimension was classified as “Excellent,” “Good,” or “Poor” based on predefined thresholds, corresponding to scores of 3, 2, and 1, respectively. The overall score for a quality dimension was calculated as the weighted sum of these scores, ranging from 1 to 3, and rounded to the nearest integer. Similarly, the overall quality of SD was determined by summing the weighted scores of all quality dimensions and rounding to the nearest integer. The final classification—"Excellent," "Good," or "Poor"—was based on scores of 3, 2, or 1, respectively. (13) The weighting scheme allowed flexibility, enabling evaluators to prioritize specific quality dimensions according to their preferences. Therefore, reasonably combining Hernadez's scoring system with the methods and metrics outlined in Chapter 4 could result in a more complete operation. Another next step following this thesis is to unify the methods and metrics from Chapter 4 with the framework suggested by Vallevik et al. in Subchapter 5.4, which considers the full process of SD generation and application. It is important to reveal that Vallevik's study is a thought-provoking research that deserves more detailed investigation than what was presented in this thesis, so those interested can refer to (10).

It is important to highlight that in this thesis, biomedical signals primarily refer to ECG signals, as most of the selected articles on this topic focus on ECG. However, one selected article examined heart sound data, which, while still considered a biomedical signal, utilized methods specific to sound data analysis for heart sound evaluation (21). This approach falls slightly outside the scope of this thesis, as the focus is on evaluating the time-series dependency in synthetic data. Another simplifying assumption made in this study is the tabular EHRs data group in the classification of synthetic data. According to Hernandez et al., biomedical signals, medical images, free-text notes from EHRs, time-series data from smart homes and living labs, activity data, and tabular data extracted from EHRs are the primary forms of synthetic health data that have been explored (13). Therefore, future research could focus on studying each type in greater detail.

As the last remark, methods and metrics that can be used to evaluate the Fairness and Carbon Footprint aspects of synthetic data need to be studied more extensively than the current situation. Moreover, the methods and metrics that evaluate Realism, Utility, and Privacy lack attention to time-series data, as specified in Section 6.4, so additional research is needed to address this issue as well.

8. CONCLUSION

This literature review discovered five dimensions of evaluating synthetic health data quality, concentrating on synthetic ECG signals and EHRs: Realism, Utility, Privacy, Fairness, and Carbon Footprint. Besides that, this thesis explored a framework, proposed by Vallevik and their partners, that guides the quality assurance process by explaining the considerations for each quality dimension at every stage. This study provided a comprehensive and clear understanding of methods one can use when assessing a quality dimension, except for Fairness and Carbon Footprint. The content of this study was compiled from various contemporary research sources to ensure thoroughness and inclusivity while harmonizing the evaluation strategies presented in those studies. Consequently, this research contributes to the development of generative models for synthetic health data, fostering stakeholders' confidence in its utilization.

To ensure that the findings of this study are more concrete and relevant to real-world applications, several future research directions can be explored. First, the examination of Fairness and Carbon Footprint should be conducted as thoroughly as other quality dimensions analyzed in this thesis. Moreover, studies on synthetic biomedical signals such as ECG and EEG should also be expanded. Second, the evaluation methods for each quality dimension should collectively provide a thorough assessment, which could be a simple classification such as "good" or "bad" or a more detailed grading system like "excellent," "good," and "bad." Thus, further research should explore what type of grading system would be most suitable for different circumstances. Third, when applying Vallevik's framework to any synthetic dataset, it should clearly indicate, at each stage, which evaluation methods are applicable to assess each quality dimension. Based on these evaluations, a grade should be assigned to each quality dimension at every stage. Ultimately, the framework should provide a definitive "Yes" or "No" answer to whether the synthetic data successfully fulfills its intended purpose, accompanied by relevant remarks or considerations. Besides these three directions, solutions should be explored to tackle the limitations of the evaluation methods identified in Chapter 6 of this study.

The integration of synthetic health data into healthcare offers significant potential for advancing medical research, enhancing patient care, and preserving privacy. By generating realistic datasets that mirror patient information, synthetic data facilitates the development and testing of new treatments, medical devices, and diagnostic tools without compromising individual confidentiality. This approach addresses data scarcity and privacy concerns, enabling researchers to conduct comprehensive studies while adhering to ethical standards (3, 4). However, challenges persist in ensuring the quality of synthetic data, as biases present in original datasets can be maintained, leading to skewed outcomes. Moreover, the absence of standardized benchmarks for evaluating synthetic data quality necessitates the development of robust assessment frameworks (1). Ethical considerations are paramount, particularly in maintaining fairness and preventing unwarranted use (22). As artificial intelligence continues to evolve, interdisciplinary collaboration, for example, between experts in healthcare, artificial intelligence, ethics, and policy-making, will be essential to navigating these complexities, ensuring that synthetic health data contributes positively to healthcare innovation while upholding ethical principles.

REFERENCES

1. Murtaza H, Ahmed M, Khan NF, Murtaza G, Zafar S, Bano A. Synthetic data generation: State of the art in health care domain. Comput Sci Rev [Internet]. 2023 May [cited 2025 Jan 6];48:100546. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1574013723000138

2. Reiner Benaim A, Almog R, Gorelik Y, Hochberg I, Nassar L, Mashiach T, et al. Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. JMIR Med Inform [Internet]. 2020 Feb 20 [cited 2024 Dec 31];8(2):e16492. Available from: http://medinform.jmir.org/2020/2/e16492/

3. Lam J, Boyd A, Linacre R, Blackburn R, Harron K. Generating synthetic identifiers to support development and evaluation of data linkage methods. Int J Popul Data Sci [Internet]. 2024 Jul 1 [cited 2024 Dec 31];9(1). Available from: https://ijpds.org/article/view/2389

4. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. Npj Digit Med [Internet]. 2023 Oct 9 [cited 2024 Dec 31];6(1):186. Available from: https://www.nature.com/articles/s41746-023-00927-3

5. Achterberg JL, Haas MR, Spruit MR. On the evaluation of synthetic longitudinal electronic health records. BMC Med Res Methodol. 2024 Aug 14 [cited 2025 Jan 6];24(1):181. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-024-02304-4

6. Rankin D, Black M, Bond R, Wallace J, Mulvenna M, Epelde G. Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. JMIR Med Inform [Internet]. 2020 Jul 20 [cited 2025 Jan 6];8(7):e18910. Available from: http://medinform.jmir.org/2020/7/e18910/

7. Eigenschink P, Reutterer T, Vamosi S, Vamosi R, Sun C, Kalcher K. Deep Generative Models for Synthetic Data: A Survey. IEEE Access [Internet]. 2023 [cited 2025 Jan 3];11:47304–20. Available from: https://ieeexplore.ieee.org/document/10122524/

8. Emam KE, Mosquera L, Zheng C. Optimizing the synthesis of clinical trial data using sequential trees. J Am Med Inform Assoc. 2021 Jan 15;28(1):3–13.

9. Lacasa M, Prados F, Alegre J, Casas-Roma J. A synthetic data generation system for myalgic encephalomyelitis/chronic fatigue syndrome questionnaires. Sci Rep. 2023 Aug 31;13(1):14256.

10. Vallevik VB, Babic A, Marshall SE, Elvatun S, Brøgger HMB, Alagaratnam S, et al. Can I trust my fake data – A comprehensive quality assessment framework for synthetic tabular data in healthcare. Int J Med Inf [Internet]. 2024 May [cited 2024 Dec 31];185:105413. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1386505624000765

11. Karttunen P. LARGE LANGUAGE MODELS IN HEALTHCARE DECISION SUPPORT. Faculty of Medicine and Health Technology, Tampere University. 2023 August [cited 2024 December 15]. Page 5. Available from: https://urn.fi/URN:NBN:fi:tuni-202308117563

12. David Foster. Generative Deep Learning, 2nd Edition. O'Reilly Media, Inc; 2023 [cited 2024 December 29]. Available from: https://learning.oreilly.com/library/view/generative-deep-learning/9781098134174/

13. Hernadez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic Tabular Data Evaluation in the Health Domain Covering Resemblance, Utility, and Privacy Dimensions. Methods Inf Med [Internet]. 2023 Jun [cited 2024 Dec 31];62(S 01):e19–38. Available from: http://www.thieme-connect.de/DOI/DOI?10.1055/s-0042-1760247

14. El Kababji S, Mitsakakis N, Fang X, Beltran-Bless AA, Pond G, Vandermeer L, et al. Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets. JCO Clin Cancer Inform [Internet]. 2023 Sep [cited 2025 Jan 6];(7):e2300116. Available from: https://ascopubs.org/doi/10.1200/CCI.23.00116

15. Monachino G, Zanchi B, Fiorillo L, Conte G, Auricchio A, Tzovara A, et al. Deep Generative Models: The winning key for large and easily accessible ECG datasets? Comput Biol Med [Internet]. 2023 Dec [cited 2024 Dec 31];167:107655. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0010482523011204

16. D'Amico S, Dall'Olio D, Sala C, Dall'Olio L, Sauta E, Zampini M, et al. Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology. JCO Clin Cancer Inform [Internet]. 2023 Jun [cited 2024 Dec 31];(7):e2300021. Available from: https://ascopubs.org/doi/10.1200/CCI.23.00021

17. Budu E, Etminani K, Soliman A, Rögnvaldsson T. Evaluation of synthetic electronic health records: A systematic review and experimental assessment. Neurocomputing [Internet]. 2024 Oct [cited 2025 Jan 6];603:128253. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0925231224010245

18. Kim J, Choo H, Shin SY, Song KD. Synthesis and quality assessment of combined time-series and static medical data using a real-world time-series generative adversarial network. Sci Rep [Internet]. 2024 Aug 17 [cited 2025 Jan 6];14(1):19064. Available from: https://www.nature.com/articles/s41598-024-69812-7

19. Snell KIE, Levis B, Damen JAA, Dhiman P, Debray TPA, Hooft L, et al. Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). BMJ [Internet]. 2023 May 3 [cited 2025 Mar 6];e073538. Available from: https://www.bmj.com/lookup/doi/10.1136/bmj-2022-073538

20. Ryu KH, Baek HJ, Gho SM, Ryu K, Kim DH, Park SE, et al. Validation of Deep Learning-Based Artifact Correction on Synthetic FLAIR Images in a Different Scanning Environment. J Clin Med [Internet]. 2020 Jan 29 [cited 2025 Jan 6];9(2):364. Available from: https://www.mdpi.com/2077-0383/9/2/364

21. Narváez P, Percybrooks WS. Synthesis of Normal Heart Sounds Using Generative Adversarial Networks and Empirical Wavelet Transform. Appl Sci [Internet]. 2020 Oct 8 [cited 2025 Jan 6];10(19):7003. Available from: https://www.mdpi.com/2076-3417/10/19/7003

22. Hanna MG, Pantanowitz L, Jackson B, Palmer O, Visweswaran S, Pantanowitz J, et al. Ethical and Bias Considerations in Artificial Intelligence/Machine Learning. Mod Pathol [Internet]. 2025 Mar [cited 2025 Mar 22];38(3):100686. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0893395224002667

23. Andor. Accessed on 2024 November 10. Available at: https://andor.tuni.fi/discovery/search?vid=358FIN_TAMPO:VU1&lang=en

24. PubMed. Accessed on 2024 December 15. Available at: https://pubmed.ncbi.nlm.nih.gov/

25. Zotero | Your personal research assistant. Accessed on 2024 December 13. Available at: https://www.zotero.org/