Journal of Informatics and Web Engineering

Vol. 5 No. 2 (June 2026) eISSN: 2821-370X

Synthetic Data Generation for Healthcare and Wellness: Methods, Applications, and Future Directions

Sellappan Palaniappan[1], Kasthuri Subaramaniam[2*], Oras Baker[3], Bui Ngoc Dung[4], Sumit Dhariwal[5]**

[1]Corporate Office, HELP University, No. 15, Jalan Sri Semantan 1, Bukit Damansara 50490 Kuala Lumpur, Malaysia.
[2]Department of Decision Science, Faculty of Business and Economics, Universiti Malaya, 50603 Kuala Lumpur, Malaysia.
[3]Faculty of Computing and Emerging Technology, Ravensbourne University London, 6 Penrose Way Greenwich Peninsula London SE10 0EW, United Kingdom.
[4]University of Transport and Communications, No.3 Cau Giay Street, Lang Thuong Ward, Dong Da District, Hanoi, Vietnam.
[5]Centre for AI, Madhav Institute of Technology & Science, Gola ka Mandir, Gwalior, Madhya Pradesh - 474005, India.
**corresponding author: (s_kasthuri@um.edu.my; ORCiD: 0000-0003-0704-923X)*
***corresponding author: (O.alhassani@rave.ac.uk; ORCiD: 0000-0002-0958-4861)*

Abstract - Artificial intelligence in healthcare relies heavily upon the availability of high-quality datasets, but very rigid privacy regimes, institutional silos, ethical issues, and heterogeneous data still serve to limit the availability of real-world clinical data. To combat these limitations, you focus on synthetic data generation as a privacy-preserving and scalable alternative to traditional data generation for healthcare and wellness studies. More relevant is a work on a broad and clear framework of synthesis methods based on statistical modelling, rule-based generation and domain-specific clinical logic, in contrast to recent studies that mainly target sophisticated methods for the deep learning architectures. The paper reviews the major synthetic data generation approaches like statistical distributions, machine learning techniques, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), diffusion models, and hybrid methods. Furthermore, the paper demonstrates how physiological constraints, time structures, prevalence modelling, and clinically meaningful associations could be incorporated into synthetic datasets in a range of healthcare fields including clinical risk prediction, wearable analytics, mental health text generation, genomics, epidemiological modeling and medical imaging. To ensure reproducibility and accessibility for both research and practice, practical Python-based examples along with domain-aware probabilistic models are provided. Further, it discusses evaluation approaches for evaluating statistical accuracy, downstream utility, and privacy in the process, and stresses their trade-offs on realism, technical efficiency, and disclosure potential. Future avenues for research are also presented, (e.g., digital twins, multimodal patient simulation, long-term disease progression modelling, differentially private generative systems). This work would be a theoretical basis as well as a practical guide for researchers, clinicians, and educators striving to create safe, transparent, and trustworthy synthetic healthcare data for advance of healthcare artificial intelligence.

Keywords— Synthetic Data, Healthcare Artificial Intelligence, Privacy Preservation, Generative Models, Data Augmentation

Received: 23 January 2026; Accepted: 24 March 2026; Published: 16 June 2026

This is an open access article under the CC BY-NC-ND 4.0 license.

1. INTRODUCTION

Healthcare and wellness analytics have evolved in a disruptive manner thanks to the rapid evolution of Artificial Intelligence (AI), which can support the rapid evolution of new fields such as early diagnosis, clinical risk prediction, prediction of epidemiological variables, wearable and bio-sensor analytics, genetic-risk modelling and personalised therapies. Yet, healthcare is still severely constrained from high-quality healthcare data access by strict privacy regulations, institutional silos, ethical issues, non-standard data quality, patient confidentiality, and data protection. As emphasized by multiple reviews, data limitations and privacy restrictions remain key challenges for AI research in healthcare [1], [2], [3], [4]. Synthetic data, synthetic datasets that lack identifiable patient characteristics and still maintain statistical features of real data, represent a suitable approach to overcome most of these challenges. Current works report that synthetic data may assist with privacy preservation, algorithm writing, reproducibility, simulation, and cross–institutional data sharing [5], [6], [7]. Its use has increased in high-risk, narrow or heavily regulated areas like genomics, imaging, mental health, paediatrics and rare diseases [8], [9].

Despite these advancements, constructing realistic synthetic healthcare data is intrinsically difficult. Real clinical data are exceedingly complex, with multimodal structure, highly correlated features, disease progression patterns, physiological constraints with seasonality, and nonlinear interactions. Healthcare synthetic data generation requires unique attention to modelling that is different from general-purpose synthetic data generation. The systematic reviews describe the challenges faced in generating synthetic data [10], [11]. Healthcare organizations, universities, and research institutions increasingly require datasets for developing and evaluating machine learning models, training clinicians, validating algorithms in privacy-sensitive settings, and quality assurance. They are hindered because real clinical datasets cannot be freely shared, require complex approvals, contain missingness and noise, and often underrepresent minority or rare disease populations. The use of synthetic data gives a controlled, privacy-safe alternative allowing for experimentation and minimizing legal and ethical risk [5], [6]. Even though deep generative models (GANs, VAEs, diffusion models, and LLMs) provide high fidelity, interpretable statistical and rule-based methods are important, especially for clinical education, early research, prototyping, or transparent auditing [9], [12].

However, there are remaining significant gaps in the current literature. Most attention on high-complexity generative models relies on real data for training, while less attention is given to transparent and interpretable rule-based statistical generators that play an important role in pedagogy and early-stage prototyping [8], [9], [12]. Furthermore, very few papers offer domain-specific recommendations across clinical, wearable, genomic, epidemiological, and mental health contexts, nor is there a unified framework explaining design principles, variable distributions, correlations, and label-generation strategies rooted in medical logic. To meet this requirement, we propose a robust and statistically informed foundation to provide synthetic datasets specifically designed for healthcare and wellness. Contrasting the previous work, which is focused mainly on deep learning–based generators, we highlight interpretable statistical models, medical domain logic and workflows that are practical, and applicable to scenarios without real patient data. This study is guided by the inquiry of how interpretable statistical and domain-based methods can be used to generate realistic synthetic datasets for healthcare and wellness applications. Consequently, we investigate which statistical distributions, correlations, and generative rules best model real-world clinical, physiological, genomic, textual, and epidemiological patterns, and how synthetic data can be tailored to specific healthcare domains. Our goals in this research are to offer a literature-based summary of synthetic data generation focusing on statistical and interpretable methods to build a unified design framework that explains variable selection, distribution modelling, correlation structures, class prevalence, and label-generation mechanics and to provide domain-specific synthetic dataset templates and examples. Finally, a summary of the current limitations and future directions for synthetic data research, including multimodal patient simulation and digital twin ecosystems, with the endpoint of making this paper serve both as a theoretical primer and a guidebook for researchers, clinicians, and educators alike.

This study is organized as a methodological tutorial and conceptual framework paper. Instead of pitching a novel generative technique or comparing experimental models to actual practices in the field, the aim is to consolidate the statistical groundwork, modelling logic specific to the domain at hand, and general principles of reproducible design for producing synthetic healthcare datasets. Thus, the paper provides an organized structure of framework and practical generation templates based on statistical modelling and clinical domain knowledge.

The unified synthetic data design framework shown in Figure 1, is a workflow to produce synthetic datasets of high standard with a balance of realism, analytical usefulness, and privacy protection. There are different problem definitions with the application domain, data modality, intended use, and privacy requirements. Then, dataset design defines the variables, data types, distributions as well as the correlation structure as they represent the intrinsic data features. In the statistical model setup stage, probability distributions, regression relationships, and domain-specific

constraints are written to form the procedures for data creation and generation. To that end, synthetic data generation involves generation of datasets, including sampling procedures, with consistent correlations and label relationships. It is subsequently assessed for fidelity (statistical similarity to real data), utility (use in analytical or machine learning tasks), and privacy (protection against information disclosure). Finally, a stage of calibration and iteration updates model parameters and distributions based on evaluation results up to the point where the dataset is of a suitable quality, after which the final synthetic dataset is released for use in research, experimentation or system development.

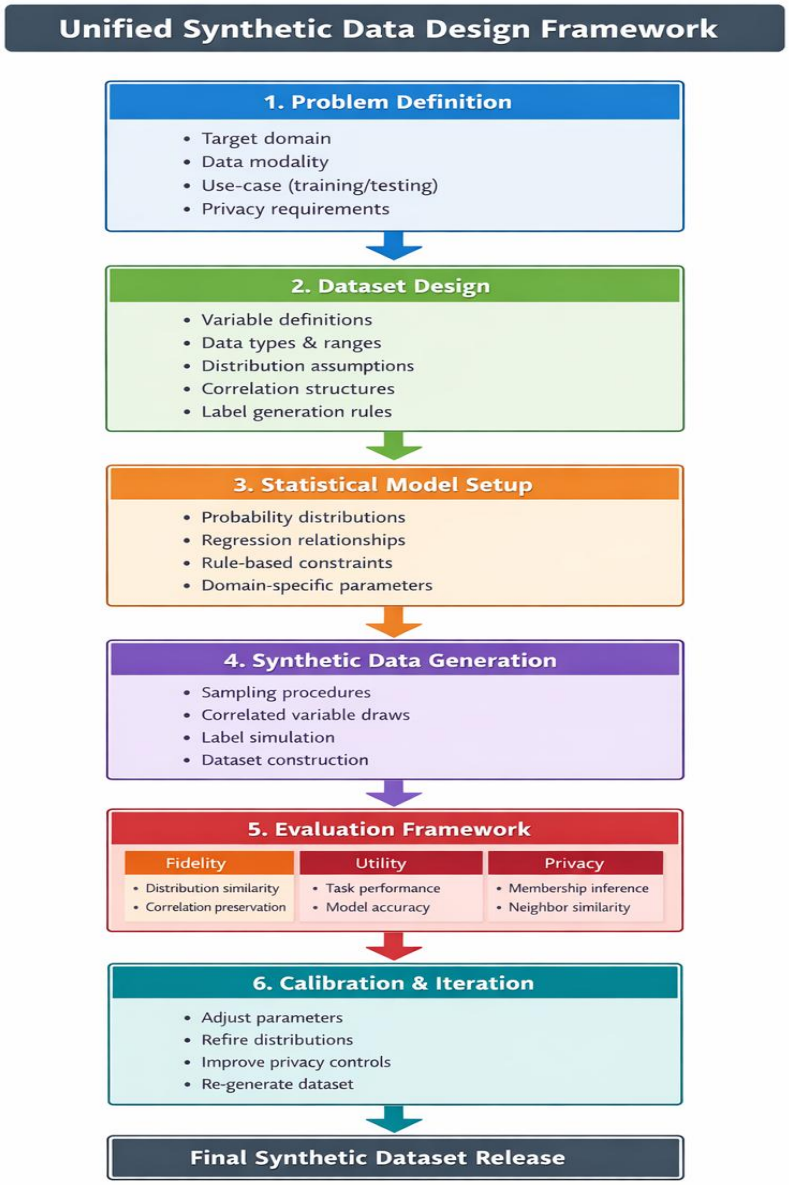

Figure 1. Unified Synthetic Dataset Design Framework

2. LITERATURE REVIEW

The digitization of healthcare services has generated large data sets, but such databases are heavily restricted access-wise by legal, regulatory and ethical considerations. Since normal anonymization methods are often insufficient to avoid re-identification attempts, synthetic data became a key avenue of privacy protection in high-dimensional medical datasets [8], [13]. Early works showed that the synthetic datasets could simulate the population statistics, with no linkage-based re-identification to occur [5], [6]. This capability is essential especially for the areas where available data are limited, such as rare diseases, paediatrics and genomics [1], [4], [7]. In a comprehensive overview this year,

Pezoulas et al. [1] observed that while deep learning techniques used in several recent studies make up more than 70% of generators, current issues with transparency, reproducibility and interpretability at the domain level remain serious challenges. Such challenges indicate that while statistical generators, uniquely positioned to resolve these challenges, are underexplored, they remain of tremendous potential utility.

Types of synthetic data generation methods can be divided into statistical and rule-based approaches, machine learning models and deep generative techniques respectively. Methods such as sampling from known distributions, generalized linear models, Bayesian mixture models and copula-based generators are desirable for their interpretability and transparency. Miletic et al. [3] and Murtaza et al. [7] point to advantages of these methods for pedagogical purposes, algorithm prototyping, and initial studies where data retrieval is limited. The success of Bayesian mixture models and Gaussian mixture-based virtual population generation in cardiology and pharmacological simulation for interpretable relationship retention is substantial based on the above-mentioned studies [14], [15], [16]. However, the rule-based technique works such as Synthea, containing medical code and disease progression path, can deliver clinically relevant and measurable accuracy and audit-prepared as result in simulation but are limited on variance [4]. Recent works have proposed various applications of AI in healthcare and health, the spectrum includes melanoma detection [17], pandemic prediction [18], predicting welfare models based on cognition [19], and co-created disease prediction models [20]. Collectively, these results emphasize the potential impact of balanced data generation methods on privacy and utility concerns.

Machine learning supports generalization of statistical methodologies due to a consideration of the conditional structures of interrelationships between variables which achieve a trade-off between interpretability and complexity. Among methods of this are three virtual population generators, regression informed simulators and conditional probability models. Lu et al. [21] discusses pipeline-workflows on dataset comparison pipelines that ease the optimisation of such generators to achieve the highest levels of fidelity and diversity, and some extend their performance in clinical trial simulation to the simulations by conditional modelling [13]. On the contrary, the use of deep learning in synthetic data Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), diffusion models, and transformer-based generators are a significant growth for synthetic data utility. Nevertheless, GAN networks such as MedGAN, CTGAN, and TTS-GAN are widely used in imaging, tabular and time series data and, as a general network for imaging, graph output, table and time series generation, they are extremely popular, but they suffer from major issues such as mode collapse, training instability and privacy leakage during training [22]. Moreover, due to the need to detect and address such issues as mode collapse, the methodologies must consider the privacy implications. The diffusion models have recently attracted much emphasis as a reliable substitute for medical imaging, and large language models have found increasing deployment to generate clinical text, which, however, raises concerns about the hallucination potential and true information of the output [4].

These approaches are implemented to varying degrees in health services. Tabular clinical data are driven by CTGAN and Gaussian Copula models, however, rule-based and regression-based generators are still widely implemented in education and simulating applications [5], [16]. Tabular synthetic data support for early-warning score model building, risk prediction prototyping, and fairness computation has been reported [11], [12]. On the other hand, the synthetic imaging research mostly uses deep learning for style transfer support, lesion augmentation, modality enhancement, etc. Time series data for physiological and wearable applications is generated using an autoregressive model and GANs based approach that considers circadian rhythms and noise characteristics for purposes such as arrhythmia classification. High dimensional genomics and omics data is generated, GAN-based and VAE-based approaches enable generating gene expression profiles, the biological plausibility and overfitting remain significant uncertainties. Epidemiological modelling often utilizes copula models, autoregressive models, and hybrid climate-infection simulators, for outbreak prediction and hotspot detection.

Notwithstanding the progress made, some important limitations have been pointed out from the literature. There is no established evaluation technique, which hinders comparison of the synthetic data quality in studies [10], [11]. Also, modelling rare diseases is still an enduring challenge because deep models that are trained on real patient data may leak privacy [12]. The area also faces challenges in modelling multimodal dependencies and uncertainty about synthetic data clinical efficacy for application as decision support in the real world. These limitations highlight a notable lack of awareness in the current literature. However, while extensive literature reviews exist on GANs, VAEs, and modern generative architectures, little work is dedicated to interpretable, auditable, rule-based synthetic data frameworks directly addressing healthcare environments. The current work fills a gap between high-level reviews and a technically complex approach of deep generative models by consolidating statistical techniques and emphasizing domain specific generative logic, thereby making synthetic data more available to researchers and clinicians.

3. PRINCIPLES OF DESIGNING REALISTIC HEALTHCARE SYNTHETIC DATA

It is necessary to develop high-quality synthetic datasets for medical and wellness applications by systematically exploring the characteristics, the science, and the roles of statistical modelling and its applications in the biomedical field. Unlike generic synthetic observations, healthcare data contains physiological limitations, disease mechanisms, timing, and interlocking clinical variables. Badly designed synthetic data sets contribute to the production of inaccurate features, associations, and biologically inconceivable outcomes, as summarized in various reviews [1], [7], [16]. Thus, the building of realistic, usable synthetic datasets need to follow a set of principled guidelines which is supported in the reality-based biomedicine for it to be able to be generated.

One of the fundamental principles is a correct definition of their parameters and range based on applicable clinical best practice. Physiological parameters (resting heart rate, blood sugar, systolic blood pressure, sleep durations, and heart rate variability: Heart Rate Variability [HRV]) provide well defined range data in clinical medicine. Not grounding variables in these norms allows the production of data very different from, or implausible to value, which makes it less realistic and less useful. Numerous studies on synthetic patient data stress the relevance of medically correct boundaries and distributions to mitigate these hazards [3], [5], [23]. Clipping, truncated distributions, or domain-informed sampling can ensure the sample complies with clinical expectations (e.g. through log-normal distributions in the case of skewed biomarkers).

Even if healthcare data exists as independent variables, they often engage well-defined physiological and disease pathways. High BMI is related to high glucose and triglycerides, HRV decreases as stress increases, and age is linked with cognitive decline. Reviews [11], [16], [21] point out that one of the major limitations of a naïve synthetic dataset is the destruction of these dependencies and subsequently derailing machine learning model training. To maintain these relationships, the methods of multivariate normal sampling with well-defined covariance matrices, regression-informed generation, copula-based modelling, and conditional rule-based approaches are employed. These methods make certain synthetic data patterns look just like real clinical data. For instance, regression-informed generation could define a relationship (1), such as:

$$Glucose = 85 + 0.7 \cdot BMI + \varepsilon \quad (1)$$

where BMI represents body mass index and $\varepsilon \sim N(0, \sigma^2)$ represents random noise capturing physiological variability.

This representation demonstrates the ways that certain biomarkers are related in order to give physiological plausibility to predictive models by realistic training. Integrating the prevalence of disease and the class imbalance is as important, since the majority of health outcomes have imbalances. For instance, sepsis, being hospitalized for heart failure or having positive cancer biopsies have special prevalence rates that, when improperly presented, cause models to overestimate precision by artificially class balancing [24]. Synthetic datasets should therefore model realistic prevalence rates, or deliberately rebalance designs, depending on the purpose of the study. Logistic transformation of risk scores, threshold rules, percentile cutoffs and hybrid models are some examples of label generation methods to retain clinically meaningful prevalence. Logistic transformation of risk score can be expressed as:

$$p(y = 1) = 1 / (1 + e\text{^}(-s)) \quad (2)$$

This distribution should represent real clinical populations and not generate artificial balance, thereby contributing to strong model validity and potential for translational purposes. Adding clinical logic to label generation is also a prerequisite for the scientific meaningfulness of labels. Realistic datasets incorporate literature-derived associations in addition to probability labels (e.g., a higher risk of hypertension associated with age, BMI (body mass index), or respiratory distress associated with low oxygen saturation). Much the same way as computational models of virtual clinical trial populations [14] and synthetic EHR engines do. Clinical logic can be translated through weighted linear combinations, rule-based thresholds, decision trees, probabilistic functions, or disease progression modules in longitudinally distributed data, which provides interpretable data. When we consider that healthcare data often presents temporal structures, we find the temporal, seasonal, and behavioral patterns in the data to be of significance too. Chronological data may include circadian rhythms, weekly cycles, seasonal variation, and autocorrelation. Without encoding of this structure, fidelity is lost and value in forecasting tasks is lost. The most popular techniques to model these dynamics are autoregressive models, seasonal patterns using sinusoidal patterns integrated with noise and conditional generative models using lag characteristics. A seasonal trend for example could look like:

$$x\_t = \mu + A \cdot sin(2\pi t / 365) + \varepsilon\_t \quad (3)$$

More complex temporal signals can be created by GANs or diffusion models [22]. In healthcare regulatory and ethical guidelines, interpretability and auditability are important aspects to consider. Transparency has also been emphasized in the literature [6], [12], [14], emphasizing that, while deep generative models create data that is quite realistic, they are usually a black box that hides the generation process. Conversely, interpretable (statistical) techniques allow full disclosure of variable definitions, auditability of generation rules, and controlled experiments that is especially useful for clinicians and educators. Maintaining privacy without compromising utility requires restricting one-to-one mapping of real patient records, having good heterogeneity in records, and avoiding nearest-neighbour similarity. Several reports suggest statistical synthetic data that has not been cross trained on actual patient datasets are the risk factors of re-identification is lower than those for deep synthetic models fed with sensitive datasets [3], [6], [12]. Finally, the heterogeneity of healthcare information data types suggests that no one synthetic approach is sufficient. This necessitates a modular and domain-specific design framework for the hybridization of domain-specific logic with statistical systems [1], [4], [7], [16]. The practical approach requires the establishment of key domain variables, appropriate statistical distribution, correlation/rule encoding, clinical label generation with clinical logic, and statistical fidelity validation. We conclude that synthetic healthcare datasets must offer biologically viable ranges, clinically salient interactions, representable prevalence levels, flexible timing, domain-specific label interpretation, and interpretability and privacy. These principles guide how synthetic data can be used to enable accurate model development, relevant benchmarking, responsible education, and safe experimentation, thus supporting trustworthy healthcare AI.

4. STATISTICAL FOUNDATIONS FOR HEALTHCARE SYNTHETIC DATA GENERATION

Laying down a foundation for statistical modelling is essential to construct realistic synthetic healthcare and well-being data sets. Some of the relatively straightforward, auditable and easily controllable methods for creating artificial health data are statistical methods, and have a particular use case under regulations, for example medicine [1], [3], [7]. Although deep generative models may obfuscate underlying mechanisms or may result in memorization of sensitive patterns, statistical foundations make explicit the specifying of distributions, correlations, constraints, and noise properties. In this section, the overall statistical ideas and approaches adopted to produce interpretable synthetic data within the healthcare area are mentioned. As healthcare data are heterogeneous, comprising continuous, discrete, symmetric and skewed characteristics, the selection of suitable statistical distribution for each of the variables is especially crucial in obtaining believable statistics. Indeed, several physiological parameters, including heart rate, systolic and diastolic blood pressure, high-density lipoprotein, and heart rate variability usually assume normal distributions within healthy populations. Gaussian models are the basis of every function, from clinical reference ranges to parametric statistical testing. Clipping extreme tails prevents physiologically unreasonable estimates when simulating these variables. Normal sampling for vitals and lab tests is one of the available statistical approaches for synthetic data, not least due to its interpretability and correspondence to population-level behaviour [7], [23]. Conversely, skewed clinical variables that have no negative value, including glucose, triglycerides, creatinine, hospital costs, and activity metrics, are often best captured by log-normal or gamma distributions. Because of their positive skew, time-series and wearable sensor studies often show activity and metabolic markers represented as these distributions.

Binary or categorical health outcomes naturally align with discrete distributions such as Bernoulli, binomial, and multinomial distributions. These are applicable for disease presence, medication adherence, genetic alleles, or symptom categories. Genomic synthetic datasets, particularly SNP simulations, routinely use binomial sampling to represent Hardy–Weinberg equilibrium allele frequencies. Furthermore, count-valued clinical or epidemiological variables, such as the number of ER visits, weekly infection counts, or symptom counts, often require Poisson-family models, with negative binomial models handling overdispersion common in infectious disease datasets and health-service utilization [6]. Proportions and probabilities, such as stress probability scores, sleep efficiency, and patient satisfaction ratings, are bounded between 0 and 1, making the beta distribution the natural choice. Studies on synthetic EHR risk prediction often use beta distributions as intermediate latent variables before converting them into categorical labels.

Healthcare variables are interdependent by nature, and unrealistic synthetic datasets often fail at the level of relationships among variables rather than individual variable distributions. Current reviews frequently highlight the importance of maintaining covariance structures to preserve clinical fidelity [11], [21]. When variables are

approximately Gaussian, multivariate normal sampling allows explicit specification of covariance matrices, defined as:

$$X \sim N(\mu, \Sigma) \text{ (4)}$$

This is useful for generating correlated lab panels, vital-sign groupings, and cognitive scores. Regression-guided generation is the other approach which provides a rigid structure for dependencies. A relationship could, for instance, be described as follows:

$$Glucose = 80 + 0.6 \cdot BMI + 0.3 \cdot age + \varepsilon \text{ (5)}$$

The method is widely used in virtual patient population models and synthetic clinical trial simulation [14]. In addition, copulas are especially well-suited for modelling non-Gaussian dependencies, where the idea is to decompose a marginal distribution from a joint structure. Both Gaussian and vine copulas have come into Favor for tabular synthetic data generation and can outperform simple correlation matrices for variables that are skewed or categorical [1], [25]. Importantly, clinically meaningful labels must be generated as one of the most criticized premises of naive synthetic dataset is that the labels can be random or uncorrelated with input variables [26]. Good label generating should encode rules based on medicine. One common approach is to generate a composite risk score, and then transform it into a probability:

$$p = 1 / (1 + e\text{^}(-score)) \text{ (6)}$$

It imitates real-life clinical risk calculators such as cardiovascular models to some extent. Threshold-based label assignment is also intuitive and based on clinical guidelines, such as flagging diabetes when the blood glucose is over 126 mg/dL. In order to avoid overly deterministic datasets, label probabilities may be composed of random noise, nonlinear transformations, or conditional interactions, using hybrid models that are based on statistical rules and medically inspired heuristics as seen in synthetic EHR engines such as SynSys and Synthea [23].

Longitudinal and time-series healthcare data modelling is important for many biomedical applications but relatively difficult to represent, and issues with temporal fidelity have been identified as one of the major shortcomings in synthetic time series health data. Daily heart rate, blood glucose, or respiratory rate are all autoregressive processes where:

$$X\_t = \phi X\_\{t - 1\} + \varepsilon\_t \text{ (7)}$$

This autocorrelation reflects biological stability and inertia. Many physiological and epidemiological variables also exhibit periodicity, modelled as:

$$X\_t = \mu + A \cdot sin(2\pi t / 365) + B \cdot cos(2\pi t / 365) + \varepsilon\_t \text{ (8)}$$

Examples include annual influenza peaks, seasonal dengue incidence, and circadian HR variation. Mixed models or hierarchical sampling allow patient-level random effects:

$$X\_i, t = \mu + u\_i + \phi X\_\{i, t - 1\} + \varepsilon\_i, t \text{ (9)}$$

The techniques generate natural within-person variation while maintaining population-level structure, which is important for digital phenotype modelling and multi-day wearable datasets.

It is critical that controlled noise and variability are included too, since synthetic datasets that are overly tidy incentivize overfitting and create an artificial performance boost. The majority of papers emphasize repeatedly that realistic stochasticity must be used to account for measurement errors at the laboratory level, sensor noise, patient differences, and environmental influence [7]. Some common noise types are Gaussian measurement error, random missingness in the data, temporal jitter of signals in wearable devices, and categorical ambiguity of the results in the dataset. It is equally relevant when dealing with missing and imperfect values: real health data is almost never complete. Missingness embeds clinical context, like ordering only when clinically indicated. Synthetic data sets must include random missing values to simulate sensor dropout, conditioned missingness where low-risk patients undergo fewer lab tests, and block missingness in time series. Simulation of missingness makes synthetic datasets more suitable for real EHR behaviour [3] as well as training ML models for field applications. Lastly, synthetic datasets need to be checked to ascertain the consistency of the statistical properties of the intended distribution or real analogues. Some significant techniques are: distribution overlap scores, correlation matrix comparison, PCA or t-SNE plot, feature importance stability in real- and synthetic-trained models. New benchmarking studies have recently focused on

comparing downstream ML performance as one potential indicator of synthetic data utility in this area [5]. Statistical foundation for realistic, interpretable, and clinically relevant generation of synthetic health data based on data is central to practical synthetic health data processing. Researchers can make synthetic data sets without exposing sensitive information by selecting realistic distributions, treating dependencies, embedding domain logic in labels, modelling dynamic temporal representations, building in noise, and implementing statistical quality.

5. DOMAIN-SPECIFIC SYNTHETIC DATASET EXAMPLES

Healthcare and wellness applications use a myriad of data methods which include tabular clinical variables, wearable sensor data, epidemiology time series, genomic variants, imaging patches, and unstructured text. Each domain has distinct statistical behaviours, physiological limitations and inter variable relationships. This section discusses a set of systematically structured synthetic dataset architectures, built on domain principles and informed by previous theoretical understandings regarding statistical modelling. The objective is not solely to show how to create synthetic healthcare datasets, but why it is the best way to model dataset from various domains. In keeping with the literature, we emphasise transparent statistical and hybrid strategies, with attention drawn less on deep generative models (GANs, VAEs): extensively covered at other places [1], [7]. The models are good examples for use in tutorial, prototype, simulation studies and to learn preliminary model.

5.1 Clinical Risk Prediction Dataset

Many clinical prediction activities, including those predicting risk associated with diabetes, readmission or cardiovascular complications, are based on tabular predictors. These datasets are combination of demographics, physiological measures, and laboratory markers that have moderate correlations and clinically meaningful thresholds.

A simulated clinical dataset may include as shown in Table 1:

Table 1. Variable Design and Distributional Choices

Variable	**Distribution**	**Domain Logic and Constraints**
Age	Uniform (30, 85)	Broad adult population coverage ; constrained between 30-85
BMI	μ=28, σ=6	Mild right skew; clipped to ensure strictly positive, biologically plausible values (e.g., > 12)
Glucose	Log-normal (μ=4.8, σ=0.25)	Highly skewed distribution typical of metabolic markers; strictly positive
Systolic BP	Normal (μ=125, σ=15)	Classical vital sign distribution; bounded realistically (e.g., 70-200)
Heart Rate	Normal (μ=75, σ=10)	Resting HR distribution
WBC Count	Normal (μ=7, σ=1.5), clipped	Clinical lab reference range

Correlations are introduced via regression-style dependencies. For example:

- Higher BMI → higher glucose
- Higher age → higher systolic BP
- WBC slightly correlated with HR (stress or inflammation patterns)

A risk score (shown in Table 2) can be computed as:

$$Score = 0.03(Age) + 0.1(BMI) + 0.02(Glucose) + \epsilon$$

Disease label assigned via Bernoulli (p), with prevalence typically ~20–30%.

Table 2. Sample Output (First 5 Rows)

Age	BMI	Glucose	Systolic BP	Heart_Rate	WBC	Risk
52.3	31.2	118.5	135.4	78.9	7.2	1
60.1	29.6	102.3	138.7	70.5	6.8	0
44.7	27.3	90.4	121.9	82.3	7.9	0
70.8	33.1	150.7	145.6	77.2	7.0	1
36.4	24.5	80.1	118.0	71.4	6.5	0

5.2 Wearable and Wellness Analytics Dataset

Wearable sensors generate continuous and often highly variable signals such as steps, heart rate, HRV, and sleep metrics. These variables exhibit temporal patterns, correlations with lifestyle, and physiological boundaries. The variable design is depicted in Table 3 and the sample output is shown in Table 4.

Table 3. Variable Design

Variable	Distribution	Domain Logic and Constraints
Heart Rate	Normal (μ=75, σ=12)	Daily average HR; bounded > 30 bpm
HRV	Normal (μ=50, σ=15)	Clipped to prevent negative values; lower values indicate physiological stress
Sleep Hours	Normal (μ=6.8, σ=1.2)	Typical adult patterns; bounded between 0 and 24 hours
Steps	Log-normal (μ=8, σ=0.4))	Highly right-skewed; strictly positive

Correlation rule:

$$stress\_score = 0.03(HR) - 0.04(HRV) + 0.01(steps/1000)$$

Table 4. Sample Output

HR	HRV	Sleep	Steps	Stress
82	45	6.3	9500	1
76	55	7.1	12000	0
88	40	6.0	8000	1
70	60	7.8	10400	0
79	42	5.9	6800	1

5.3 Mental Health Text Dataset

Text-based synthetic data is particularly useful for sentiment analysis, mood scoring, or mental health classification. Rule-based generation avoids risks associated with LLMs reproducing sensitive real-world content.

There are three sentiment categories for dataset logic:

- **Negative (−1):** fatigue, stress, anxiety
- **Neutral (0):** routine statements
- **Positive (1):** optimism, energy

Templates are combined with interchangeable phrase components.

5.4 Genomic SNP Dataset

Synthetic genomic data must respect allele frequencies, linkage patterns, and polygenic risk scoring.

For allele frequency **p**, genotype frequencies using SNP Generation with Hardy–Weinberg Equilibrium:

- 0 copies (AA): (p^2)
- 1 copy (Aa): (2pq)

- 2 copies (aa): (q^2)

Weights are applied to compute a Polygenic Risk Score (PRS):

$$PRS = \sum_{i=1}^{k} w_i \cdot SNP_i$$

5.5 Synthetic Medical Imaging (Patch-Based)

Although deep models dominate synthetic imaging research, simple rule-based patch generators can provide usable datasets for teaching CNNs.

The logic includes:

- Base image: Gaussian noise
- Lesion-like patch: circular or square bright region
- Optional segmentation mask

No sample output is shown here since it produces images, but the patch array values behave realistically for CNN input.

5.6 Epidemiological Time-Series Dataset

Epidemiological modelling requires temporal structure, climate variables, and lagged case dependence as depicted in Table 5.

Table 5. Variable Set

Variable	**Rationale**
Temperature	Seasonal sinusoidal trend
Rainfall	Gamma distribution; impacts vector populations
Humidity	Correlated with temperature
Cases	AR(1) + climate effects
Lagged Cases	Autocorrelation driver
Hotspot Label	Threshold-based

In summary, the domain-specific examples presented above demonstrate how statistical distributions, correlation structures, physiological logic, and domain expertise can come together to produce meaningful synthetic datasets across healthcare and wellness domains. These datasets are used in several applications including but not limited to clinical prediction, wearable analytics, genomics, imaging, and epidemiology. Every sample exemplifies the rationales behind the design, details the implementation, and practical Python code, so readers have an arsenal in building their own datasets to support research.

6. EVALUATING SYNTHETIC DATA FIDELITY, UTILITY, AND PRIVACY

Assessing synthetic healthcare data is an important step in deciding if the generated datasets can realistically be used to guide model construction, useful for subsequent analysis, and private enough to prevent any re-identification threat. Unlike real data which can be evaluated through clinical ground truth, the synthetic datasets call for a multi-dimensional evaluation framework because the "correctness" of synthetic data depends on statistical similarity, model performance similarity, and privacy protection, all simultaneously. An increasing number of papers have shown that synthetic data needs to achieve a high compromise between the three pillars such as fidelity, utility, and privacy for responsible applications in healthcare AI [1], [12], [24]. It is especially challenging to maintain high fidelity without

affecting privacy, as excessively realistic synthetic data can accidentally surface the most sensitive elements. This tension highlights the importance of strong evaluation frameworks and metrics.

In the following section, we introduced (i) the fundamental concepts of evaluation, (ii) methods for assessing statistical fidelity, (iii) utility evaluation via performance of downstream models, and (iv) assessment of privacy and disclosure risk, with examples from practical contexts that apply to health and wellness datasets.

6.1 Statistical Fidelity Evaluation

Statistical fidelity measures how closely synthetic data fits the underlying structure of the real data distribution that is all without the need for record-level correspondence. High-fidelity data preserve patterns overall but not replicate individuals. There are three elements to statistical fidelity.

The first element is univariate fidelity. Table 6 shows an intuitive comparison of individual feature distributions across real vs synthetic datasets. Typical methods include:

- Histogram and kernel density comparison
- Summary statistic comparison (mean, median, variance, skewness)
- Kolmogorov–Smirnov (KS) test
- Anderson–Darling or Cramér–von Mises tests
- Jensen–Shannon Divergence (JSD) for probabilistic comparison

Table 6. A Description of Fidelity Report

Feature	**Mean (Real)**	**Mean (Syn)**	**KS Statistic**	**p-value**
Glucose	112.4	114.1	0.08	0.52
HR	76.1	76.9	0.05	0.68

In good synthetic data, KS statistics should be small, and p-values should not be significant, indicating the synthetic feature distributions are not statistically different from the real ones.

The second element is multivariate fidelity. Healthcare data has very few standalone variables which include relationships between these features, correlations, covariance structures, conditional dependencies. Thay all have an important biological significance.

The multivariate fidelity checks whether synthetic data preserves:

- Correlation matrices of the original data
- Feature–feature relationships via scatterplots
- Nonlinear dependencies (e.g., BMI–glucose relationships)
- Higher-order interactions

The metrics include:

- Correlation distance:

$$d = \| C_{\text{real}} - C_{\text{synthetic}} \|_F$$

where $\|\cdot\|_F$ is the Frobenius norm.

- Mutual information comparison
  Useful for nonlinear relationships.
- Principal Component Analysis (PCA) overlap
  Plotting PC1 vs PC2 for real and synthetic data should reveal similar structure without individual overlap.

The third element is the distributional visualization. Visual inspection is still one of the most intuitive and effective ways to evaluate the fidelity of synthetic data. Typical approaches include:

- overlay real vs. synthetic variable density plots
- compare Cumulative Distribution Functions (CDFs)
- analyse central tendencies and spread via box-and-whisker or violin plots.

Correlation heatmaps showing real and synthetic pairwise relationships provide additional support for detecting structural fidelity. These images tell researchers that clinically plausible patterns such as the expected increase in glucose with age or body mass index are preserved in synthetic data and can unveil any implausible or distorted associations which might threaten validity.

6.2 Utility Evaluation

Evaluation of utilities determines whether synthetic data can provide meaningful and reliable downstream analyses. A good synthetic dataset should enable researchers to train machine learning models whose performance closely matches that of models trained on real data; they can do exploratory analyses without encountering misleading trends, validate analytical pipelines before deploying on sensitive clinical data, and reproduce relative effect sizes observed in real-world statistical models. Utility is usually assessed at two complementary levels: model-level utility and task-level utility.

6.2.1 Model-Level Utility

Traditionally, model-level utility is assessed through training a predictive model solely on synthetic data and testing it on real-world data. This result is compared with the benchmark model trained and tested on real data. Results were interpreted using standard ML metrics which include accuracy, precision, recall, F1-score, and ROC–AUC for classification tasks; Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE) for regression; and calibration curves, alignment of feature importance rankings, and similarity of regression coefficients in linear models. A synthetic dataset is useful if the performance gap between synthetic-trained and real-trained models falls within an acceptable tolerance, indicating that the synthetic data preserves the predictive structure of the original as shown in Table 7.

Table 7. Performance Gap

Model	**Train Data**	**Test Data**	**AUC**
Logistic Regression	Real	Real	0.82
Logistic Regression	Synthetic	Real	0.78

A performance reduction of ≤ 5% between the synthetic-trained and real-trained models indicates acceptable utility.

6.2.2 Task-Level Utility

The task-level utility evaluates whether synthetic data still can support valid analytical inference even when controlled noise or approximation is used. Synthetic data would have high task-level utility if it allows reliable exploratory analysis and meaningful hypothesis making. This necessitates that the direction and magnitude of important correlations follow the actual data, and that existing clinical or epidemiological associations. For instance, the positive association between increasing BMI and rising glucose levels could still be preserved. Further, synthetic datasets should capture realistic prevalence rates of conditions or outcomes, uphold class imbalance consistent with domain knowledge, and model temporal features (e.g. seasonality, long-term trends). Equally, the representation of rare events should be accurate as omission or distortion of rare events can reduce model performance and make inferences misleading.

6.3 Privacy and Disclosure Risk Evaluation

Synthetic data is assumed to be inherently privacy-preserving but exposes disclosure risk. High-fidelity generation methods, however, particularly deep learning-powered methods such as GAN, have the potential to unknowingly learn or mimic features of actual humans, and potentially re-identify them. Privacy evaluations should first identify an adversary's assumed capabilities. The threat models can generally be divided into two main types:

- White-Box Attack: attacker has full access to the architecture, parameters, and weights of the generative model, allowing them to reverse-engineer the training data.
- Black-Box Attack: attacker has only access to the released synthetic dataset and/or API endpoint; they can only see the generated outputs.

The conventional assumption for the statistical/rule-based generative healthcare systems which we will rely on in our model is a black-box threat model. Due to statistical generators not actively learning or holding the representation of real patient samples like complex generative models, their risks of white-box weight extraction are inherently low. Therefore, results must be interpreted through a strict privacy lens which include record-level uniqueness and attribute and membership inference.

6.3.1 Record-Level Uniqueness

A basic privacy check is to confirm that no synthetic record is the same as any record in the original dataset. Moreover, distance metrics like Euclidean or Mahalanobis distance should be used to check for near-duplicates. Common metrics are the minimum distance between each synthetic record and its nearest real neighbour and the proportion of synthetic samples within a small radius ε of any real record. Many synthetic points at very low ε (near real data) demonstrate overfitting and perhaps invasion of privacy.

6.3.2 Attribute Disclosure Risk

When an adversary can precisely infer sensitive or missing characteristics of an individual by using this synthetic dataset, this is called attribute disclosure. This risk can be assessed by eliminating certain attributes from a synthetic record and predicting missing values from the remaining synthetic data. If the predictive values were close and highly reliable to the original real ones, the synthetic data set presents an additional threat of attribute disclosure.

6.3.3 Membership Inference Risk

Membership inference is the ability for an attacker to decide whether the dataset for a given user has been included in the training set used to generate the synthetic dataset. This is usually assessed by training auxiliary "shadow" models to identify behaviours suggestive of overfitting and then testing the attacker's prediction accuracy. While previous studies suggest that accuracy near 50% implies low membership risk, this threshold must be adjusted for clinical class imbalances. In healthcare datasets with rare diseases, an attacker could achieve high raw accuracy simply by predicting the majority class.

6.3.4 Differential Privacy Considerations

Some such synthetic data generation models combine Differential Privacy (DP), a formal privacy provision, which involves injecting calibrated noise in summary statistics, generation processes or sample procedures. The privacy strength is controlled by the privacy budget ε: lower ε values provide better privacy but are more intrusive and may reduce statistical precision and analytical utility. This demonstrates the trade-off that always exists between protecting privacy while providing potential utility to the data, requiring careful recalibration of the balance for different use cases. To ensure transparency, the selected ϵ value, the specific DP mechanism utilized, and the data bounds must be explicitly documented when releasing the synthetic dataset so that practitioners understand the exact statistical noise floor introduced by the privacy constraint.

6.4 Fidelity–Utility–Privacy Trade-Offs

Synthetic data must balance three competing objectives as shown in Table 8:

Table 8. Trade-Offs

Dimension	Goal	Risk if maximized
Fidelity	Realistic, clinically meaningful patterns	Privacy leakage
Utility	Effective model development and analysis	Misleading conclusions
Privacy	Prevent re-identification	Loss of fidelity; unusable datasets

A perfect balance is rarely achievable; instead, the aim is optimal compromise based on use case:

- Educational datasets → prioritize utility and interpretability
- Model prototyping → prioritize fidelity and task utility
- Public release datasets → prioritize privacy over fidelity

This compromise should be clearly recognized in healthcare, where the privacy policies like GDPR, HIPAA, and PDPA exert maximum control over how data is treated. In conclusion, an assessment framework of statistical fidelity, downstream utility and privacy protection is needed to help us assess synthetic data. Strong synthetic datasets must:

- Preserve meaningful statistical and clinical patterns
- Support analytical and modelling tasks with minimal performance degradation
- Prevent identity leakage or attribute inference

These assessments verify that synthetic data is useful, reliable, and ethical to deploy in healthcare AI research. With the explosion of synthetic data in the context of digital health, extensive assessment is necessary to ensure scientific validity and regulatory compliance.

7. DISCUSSION

It was shown that the results demonstrate that through statistical distribution-based domain expertly derived data-driven synthetic data that has been developed statistically, and that can be accurately modelled health data based on existing data may provide reasonable approximations of the structure of data without compromising patient data integrity of health and patient privacy. This ability is highly desired, given that innovation is often stifled by a regulatory environment with a plethora of privacy issues and restricted health data availability restrictions, often slowing down innovation. Synthetic data enables researchers to prototype models and design the models and analysis pipelines on which they can then build models and design analysis pipelines without relying on access to sensitive health information, and by ensuring that research is both faster and easier to develop while avoiding the delays of approval as a real-world data governance system for real-world data. This method tackles the long-standing tension between a strong dataset and an ethical requirement to maintain confidentiality. Rather than asserting the superiority of one approach, this work highlights a critical trade-off. While fully learned generative models (e.g., GANs or diffusion models) excel in statistical fidelity and high fidelity, statistical and rule-based methods offer unparalleled advantages in transparency, interpretability, and privacy. By circumventing the “black box” nature of deep learning, these auditable methods are especially crucial in highly regulated environments where real patient data cannot be utilized.

While synthetic datasets offer a safe sandbox for innovation, they cannot completely substitute the complexity of real-world clinical analysis. A primary limitation of the proposed rule-based and statistical systems is the potential oversimplification of clinical diversity, which may overlook non-linear interactions and rare co-morbidities. If models are trained on data with unrealistic assumptions, they may derive spurious associations that remain hidden upon deployment. Therefore, synthetic data must be strictly scoped to prototyping, education, and pipeline testing. It must never serve as the sole basis for clinical deployment models; practitioners are obligated to validate any synthetic-trained algorithms against real-world data and rely on domain experts to verify clinical relevance before real-world application. Ultimately, synthetic data is designed to safely complement, not replace, real clinical data

Hence, future work should emphasize the advancement of complexity and temporal depth of synthetic datasets. Noteworthy approaches are the creation of multimodal synthetic patients that incorporate structured electronic health records, and integrated with wearable sensor signals, narrative notes and genomic components. Moreover, integrated temporal and longitudinal modelling to mimic disease progression and response to treatment will better align synthetic data capabilities with the potential needs of advanced analytics. To achieve formal privacy guarantees and still deliver high fidelity, we also discussed the integration of DP into generative models. Finally, development of large-scale agent-based epidemiological models based on large populations could facilitate public health planning by emulating

population and disease transmission networks. The practical application of these results requires a disciplined approach to responsible consumption. Synthetic data should be used for prototyping and education, as well as pipeline testing, not serve as the basis for clinical deployment models solely from synthetic data. Practitioners need to validate the development of models trained on synthetic data with real-world data and have domain experts check the outputs for clinical relevance. Given frameworks that require in-depth documentation of generation assumptions and scrutiny of fidelity and utility, the research community can use synthetic data to support safe, ethical, and scalable innovation in healthcare AI.

8. CONCLUSION

Today, synthetic data is a key contributor to health and wellness AI innovation, enabling a morally acceptable alternative that balances privacy while supporting model development, education, simulation, and research. In the midst of strict data protection laws and different data ecosystems, it provides a practical solution for keeping confidentiality without sacrificing the rigor of experiments. Herein, we present a statistical driven overall theoretical approach to generate artificial datasets, tailored to meet specific challenges of health care. We outline key drivers, like privacy constraints, limited data access, and highlight basic design principles: physiologically plausible ranges, realistic correlations, clinically meaningful patterns, and accurate prevalence estimates. Our methodology employs univariate and multivariate sampling, distribution modelling, correlation models, mixture models, and regression-based dependencies to support the statistical validity and interpretability of our estimates. This application has been applied to specific domains such as clinical risk prediction, wearable analytics, mental health text modelling, genomics, medical imaging, or epidemiology. These demonstrate that the correlation of statistical and domain knowledge can generate some realistic structured and semi-structured synthetic data. In this case, they can be implemented on easily reproducible and accessible generation through standard scientific libraries (Python-based implementation) without any dedicated tools. These are three evaluative criteria, namely: statistical fidelity, downstream usefulness, and privacy risk. Synthetic datasets have certain advantages, but also some disadvantages: (b) we oversimplify clinical variability; (c) we risk implausible assumptions; (d) it is hard to model rare events or multimodal interactions.

However, synthetic data are now going into greater use in research, industry, in regulatory testing and training; thus, improving reproducibility and AI safety; it also enhances interdisciplinary cooperation. Upcoming lines of research include multimodal digital patients, longitudinal disease simulations, agent-based epidemiological models, and differentially private generative frameworks. Most significantly, synthetic data complements, not replace the real clinical data. Properly generated with clinical rigor, rigorous quality assurance, and good validation ensures that it is a potent weapon to fast-track responsible implementation of AI in medicine. Its strategic incorporation in R&D pipelines will be important if we are to create trustworthy, fair, and powerful AI for worldwide health challenges.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING STATEMENT

This research received no external funding.

AUTHOR CONTRIBUTIONS

Sellappan Palaniappan: Conceptualization, Data Curation, Methodology, Writing – Original Draft Preparation;
Kasthuri Subaramaniam: Project Administration, Supervision, Methodology, Writing – Review & Editing;
Oras Baker: Methodology, Validation, Writing – Review & Editing;
Bui Ngoc Dung: Project Administration, Writing – Review & Editing;
Sumit Dhariwal: Findings, Writing – Review & Editing.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

This research did not require IRB approval because it used publicly available secondary data with no identifying personal information.

DATA AVAILABILITY

Derived data supporting the findings of this study are available from the corresponding author on request.

REFERENCES

[1] V. C. Pezoulas, D. I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. S. Tachos, and D. I. Fotiadis, "Synthetic data generation methods in healthcare: A review on open-source tools and methods," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 2892–2910, 2024, doi: 10.1016/j.csbj.2024.07.005.

[2] H. A. Ahmed, J. A. Nepomuceno, B. Vega-Marquez, and I. A. Nepomuceno-Chamorro, "Synthetic data generation for healthcare: Exploring generative adversarial networks variants for medical tabular data," *International Journal of Data Science and Analytics*, pp. 1–16, 2025, doi: 10.1007/s41060-025-00816-w.

[3] M. Miletic, and M. Sariyar, "Synthetic data generation methods for longitudinal and time series health data," *Studies in health technology and informatics*, vol. 328, pp. 367–371, 2025, doi:10.3233/SHTI250740.

[4] S. Capuozzo, "Datasets: A trustworthy approach to generate guided synthetic biomedical image samples," in *Image Analysis and Processing – ICIAP 2025 Workshops: 23rd International Conference, Rome, Italy, September 15–19, 2025, Proceedings, Part II*, Springer Nature, pp. 400, 2026, doi: 10.1007/978-3-032-11381-8_34.

[5] M. Goyal, and Q. H. Mahmoud, "A systematic review of synthetic data generation techniques using generative AI," *Electronics*, vol. 13, no. 17, Art. no. 3509, 2024, doi: 10.3390/electronics13173509.

[6] M. Rujas, R. M. Gomez del Moral Herranz, G. Fico, and B. Merino-Barbancho, "Synthetic data generation in healthcare: A scoping review of reviews on domains, motivations, and future applications," *International Journal of Medical Informatics*, vol. 195, Art. no. 105763, 2025, doi: 10.1016/j.ijmedinf.2024.105763.

[7] H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, and A. Bano, "Synthetic data generation: State of the art in health care domain," *Computer Science Review*, vol. 48, Art. no. 100546, 2023, doi: 10.1016/j.cosrev.2023.100546.

[8] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing*, vol. 493, pp. 28–45, 2022, doi: 10.1016/j.neucom.2022.04.053.

[9] B. van Breugel, T. Liu, D. Oglic, and M. van der Schaar, "Synthetic data in biomedicine via generative artificial intelligence," *Nature Reviews Bioengineering*, vol. 2, no. 12, pp. 991–1004, 2024, doi: 10.1038/s44222-024-00245-7.

[10] A. Jadon, and S. Kumar, "Leveraging generative AI models for synthetic data generation in healthcare: Balancing research and privacy," in *Proceeding of 2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*, IEEE, pp. 1–4, 2023, doi: 10.1109/SmartNets58706.2023.10215825.

[11] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in healthcare: A narrative review," *PLOS Digital Health*, vol. 2, Art. no. e0000082, 2023, doi: 10.1371/journal.pdig.0000082.

[12] J.-F. Rajotte, R. Bergen, D. L. Buckeridge, K. El Emam, R. Ng, and E. Strome, "Synthetic data as an enabler for machine learning applications in medicine," *iScience*, vol. 25, no. 11, 2022, doi: 10.1016/j.isci.2022.105331.

[13] M. Z. Uddin, *Machine Learning and Python for Human Behavior, Emotion, and Health Status Analysis*. CRC Press, 2024, doi: 10.1201/9781003425908.

[14] Y. Zhang *et. al.*, "GAN-based one dimensional medical data augmentation," *Soft Computing – A Fusion of Foundations, Methodologies & Applications*, vol. 27, no. 15, 2023, doi: 10.1007/s00500-023-08345-z.

[15] O. Mazumder, R. Banerjee, D. Roy, S. Bhattacharya, A. Ghose, and A. Sinha, "Synthetic PPG signal generation to improve coronary artery disease classification: Study with physical model of cardiovascular system," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 2136–2146, 2022, doi: 10.1109/JBHI.2022.3147383.

[16] T. Das, Z. Wang, and J. Sun, "Twin: Personalized clinical trial digital twin generation," in *Proceeding 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 402–413, 2023, doi: 10.1145/3580305.3599534.

[17] O. Baker, Z. Ziran, M. Mecella, and K. Subaramaniam, "AI-driven melanoma detection in New Zealand: A ResNet50-based approach," *Informatics in Medicine Unlocked*, vol. 58, Art. no. 101697, 2025, doi: doi.org/10.1016/j.imu.2025.101697

[18] O. Baker, Z. Ziran, M. Mecella, K. Subaramaniam, and S. Palaniappan, "Predictive modeling for pandemic forecasting: A COVID-19 study in New Zealand and partner countries," *International Journal of Environmental Research and Public Health*, vol. 22, no. 4, Art. no. 562, 2025, doi: 10.3390/ijerph22040562.

[19] S. Palaniappan, R. Logeswaran, K. Subaramaniam, O. Baker, and B. N. Dung, "Training the brain: A machine learning approach to predicting wellbeing through intentional thought pattern modification," *Journal of Informatics and Web Engineering*, vol. 4, no. 3, pp. 64–89, 2025, doi: 10.33093/jiwe.2025.4.3.4.

[20] O. Baker, K. Subaramaniam, A. S. Shibghatullah, Z. A. Shaffiei, and A. S. S. Amir Hamzah, "A collaborative framework for disease prediction using machine learning," in *2025 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia*, pp. 650–654, 2025, doi: 10.1109/IICAIET67254.2025.11265309.

[21] C. Lu, C. K. Reddy, P. Wang, D. Nie, and Y. Ning, "Multi-label clinical time-series generation via conditional GAN," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 4, pp. 1728–1740, 2023, doi: 10.1109/TKDE.2023.3310909.

[22] R. Osuala *et al.*, "medigan: A Python library of pretrained generative models for medical image synthesis," *Journal of Medical Imaging*, vol. 10, no. 6, pp. 061403-1–061403-11, 2023, doi: 10.1117/1.JMI.10.6.061403.

[23] T. N. Arvanitis, S. White, S. Harrison, R. Chaplin, and G. Despotou, "A method for machine learning generation of realistic synthetic datasets for validating healthcare applications," *Health Informatics Journal*, vol. 28, no. 2, Art. no. 14604582221077000, 2022, doi: 10.1177/14604582221077000.

[24] S. Dey, P. Basuchowdhuri, D. Mitra, R. Augustine, S. K. Saha, and T. Chakraborti, "Blimsr: Blind degradation modelling for generating high-resolution medical images," in *Annual Conference on Medical Image Understanding and Analysis*, Cham: Springer Nature Switzerland, pp. 64–78, 2023, doi: 10.1007/978-3-031-48593-0_5.

[25] J. Shi, D. Wang, G. Tesei, and B. Norgeot, "Generating high-fidelity privacy-conscious synthetic patient data for causal effect estimation with multiple treatments," *Frontiers in Artificial Intelligence*, vol. 5, Art. no. 918813, 2022, doi: 10.3389/frai.2022.918813.

[26] P. Singhal and M. Singh, "Robust medical image prediction via adaptive reconstruction: bridging the gap in low-quality data", *Journal of Informatics and Web Engineering*, vol. 5, no. 1, pp. 1–17, Feb. 2026, doi: 10.33093/jiwe.2026.5.1.1.

BIOGRAPHIES OF AUTHORS

	**Sellappan Palaniappan** is a Professor at HELP University. His research focuses on the application of artificial intelligence, machine learning, data science, and cybersecurity in diverse domains. His research interests also include quantum physics, neuroscience, energy frequency vibration, healing and wholeness, and sustainable development goals. He can be contacted at email: sellappan.p@help.edu.my.
	**Kasthuri Subaramaniam** is a Senior Lecturer at the Department of Decision Science, Faculty of Business and Economics, Universiti Malaya. Her research interests include human-computer interaction, human personality types, augmented reality, artificial intelligence and cybersecurity. She actively serves as a reviewer for refereed journals. She may be contacted at email: s_kasthuri@um.edu.my.
	**Oras Baker** is an Associate Professor and Head of Masters in Cyber Security and Cyber Security Management at University of Ravensbourne London, UK. With 25 years of distinguished experience spanning academia, research, and industry, he specialises in Artificial Intelligence, Software Engineering, Cyber Security, Data Mining, and Machine Learning. He can be contacted at email: O.alhassani@rave.ac.uk.
	**Bui Ngoc Dung** is a Senior Lecturer in Information Technology at the University of Transport and Communications, Vietnam. He holds a Ph.D. in Informatics from Malaysia University of Science and Technology and has completed postdoctoral research at TU Wien, Austria. His research focuses on machine learning, computer vision, biomedical applications, and structural health monitoring. Dr. Dung also serves as a reviewer, technical committee member, and keynote speaker at international conferences. He can be contacted at email: dnbui@utc.edu.vn.
	**Sumit Dhariwal** is an Assistant Professor at the Centre for AI, Madhav Institute of Technology & Science (MITS-DU), India. He holds a Ph.D. from Malaysia University of Science and Technology. His research interests include image processing, computer vision, AI/ML, green technology, crop disease detection, cybersecurity, and face detection. He has published over 30 papers in reputable journals and conferences. Contact: sumitdhariwal@mitsgwalior.in.