http://dx.doi.org/10.21622/ACE.2026.06.2.2356

STRUCTURAL PREPROCESSING AND FEATURE ALIGNMENT FOR OPTIMIZING CTGAN-BASED SYNTHETIC GENERATION FROM SOUTH AFRICAN GLOBAL HEALTH OBSERVATORY DATA

Junior Vela Vela[1], Prabhakar Rontala Subramaniam[2], Japheth Kiplanga't Mursi[3] and Kolitsoe Sehapi[4]

[1] Haute École de Commerce- Kinshasa, Congo
[2] University of KwaZulu-Natal, South Africa
[3] Daystar University, Kenya
[4] University of KwaZulu-Natal, South Africa

Emails: {juniorvela@heckin.ac.cd, prabhakarr@ukzn.ac.za, jmursi@daystar.ac.ke, sehapik@ukzn.ac.za}

Received 29 July 2026 - Accepted 24 September 2026 - Published 04 October 2026

ABSTRACT

High-fidelity synthetic data generation is a critical prerequisite for addressing data scarcity and privacy constraints in global health. While Conditional Tabular Generative Adversarial Networks (CTGANs) are established in synthesizing patient-level electronic health records (EHRs), their performance on aggregate, high-variance transnational health indicators remains underexplored. This study introduces the Structural Preprocessing and Feature Alignment (SPFA) Framework, a novel, preprocessing-centric pipeline designed to optimize CTGAN performance for aggregate indicators. By evaluating the efficacy of the SPFA Framework in synthesizing the WHO Global Health Observatory data for South Africa, the research shifts the focus of generative modeling from clinical settings to the macro-level health surveillance domain.

The SPFA Framework implements a five-stage hierarchical pipeline incorporating automated type-safe null handling, per-indicator IQR-based outlier capping, dual-criterion cardinality reduction, and Log1p normalization to mitigate heteroskedasticity. This approach was evaluated through a multi-criterion fidelity assessment, integrating SDV structural quality scores, statistical descriptor proximity, visual density histograms, and Pearson correlation preservation. The SPFA-optimized CTGAN achieved acceptable-to-good statistical fidelity across all dimensions, demonstrating strong marginal distributional alignment (SDV Column Shape $\geq$ 0.85) and accurate preservation of inter-feature relationships (Pearson correlation difference < 0.02). Notably, the mean and standard deviation of synthetic values differed from real-world data by only -5.81% and +1.27%, respectively. While minor divergence was observed in the upper quartile distributions, the pipeline successfully addressed the high cardinality and skewness inherent in South African health indicators. These results validate the SPFA Framework as a robust, reproducible methodology for public health. By demonstrating that structural alignment is a primary determinant of synthesis quality, comparable in influence to hyperparameter tuning, this study provides a scalable pathway for enhancing data utility in global health surveillance and evidence-based policy modeling.

Keywords: (Synthetic Data, CTGAN, WHO Global Health Observatory, Structural Preprocessing and Feature Alignment, Health Surveillance, South Africa)

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

1. INTRODUCTION

Despite substantial global investments in health data infrastructure, significant coverage gaps persist, particularly in low-and middle-income countries (LMICs). International financial support for data and statistics rose from $586 million to $875 million between 2015 and 2022; yet these resources have not translated into comprehensive and reliable data availability [1]. Similarly, [2] highlights the challenges posed by missing information in longitudinal health datasets obtained from the World Health Organization's Global Health Observatory (GHO). Using GHO data from 100 countries for 2015-2019, the authors demonstrate that increasing levels of missingness can substantially affect the reliability of health indicators and the performance of imputation methods. Their findings indicate that imputation remains relatively robust up to approximately 50% missingness, whereas missing proportions exceeding 50% require caution, and levels above 70% can result in substantial variance shrinkage and compromised data reliability. These findings underscore the importance of systematic missing-data assessment and appropriate preprocessing when GHO data are used for machine learning and predictive modeling. Compounding these issues, [3] identify substantial data gaps in health- and well-being-related Sustainable Development Goal indicators, despite improvements in data availability. Their analysis of 58 health-related SDG and World Health Organization Triple Billion indicators demonstrates considerable variation in data availability across countries and years, highlighting limitations in the longitudinal monitoring of health outcomes at the global level.

These structural deficiencies extend far beyond technical inconvenience. They directly compromise the integrity of health analytics: models trained on sparse, incomplete, or unrepresentative data are inherently prone to systematic bias, particularly against LMIC populations that remain chronically underrepresented in global surveillance systems. This creates a critical paradox: populations with the greatest need for data-driven health interventions are precisely those least represented in the datasets used to inform them.

In response to these challenges, generative modeling has emerged as a powerful approach for data augmentation. Since their introduction by [4], Generative Adversarial Networks (GANs) have evolved from image synthesis into a dominant paradigm for generating high-fidelity synthetic tabular data [5;6]. The application of generative models in healthcare has grown rapidly, with a scoping review documenting a consistent upward trend in synthetic health data research between 2015 and 2024, peaking at 27 publications in 2025[7 P].

[5] Identify three primary use cases for synthetic health data: privacy-preserving data sharing, training data augmentation for machine learning models, and the simulation of counterfactual health scenarios. The present study focuses on the second of these, training data augmentation, specifically within the context of WHO South African health indicator data. By supplementing sparse real-world datasets with statistically consistent synthetic records, the aim is to enhance data coverage and enable more robust downstream modeling [8].

Among existing approaches, the Conditional Tabular GAN (CTGAN) has become a leading method for tabular data synthesis [8; 9]. Implemented within the Synthetic Data Vault (SDV) ecosystem [10], CTGAN provides a production-grade, metadata-aware framework well-suited to modeling complex tabular distributions and is increasingly adopted in healthcare-related synthetic data research. Nevertheless, despite rapid advances in generative modeling techniques, the literature has paid comparatively little attention to the development of rigorous, reproducible preprocessing protocols [11;12]. The present study seeks to address this methodological gap by examining both the technical and practical implications of high-quality preprocessing in synthetic health data generation.

1.1. RESEARCH CONTRIBUTIONS

This study contributes to the growing body of literature on synthetic health data and AI applications in healthcare through several interrelated methodological and empirical advancements.

First, it introduces a systematic, well-documented, and fully reproducible preprocessing pipeline specifically tailored to GHO tabular data. The pipeline incorporates type-safe null

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

handling, per-indicator interquartile range (IQR) outlier capping, log1p normalization, and dual-criterion cardinality reduction. Designed as a reusable research artifact, it offers a practical foundation for other scholars working with global health indicator datasets.

Second, the study presents a novel application of CTGAN to aggregate indicator data from the WHO Global Health Observatory. This extends existing research beyond patient-level electronic health records (EHRs) into the comparatively underexplored domain of international health surveillance datasets.

Third, it implements a multi-criterion synthetic data fidelity evaluation framework that integrates SDV structural quality scores, statistical descriptor comparisons, overlaid distributional visualizations, categorical frequency fidelity assessments, and temporal correlation preservation metrics. This comprehensive approach enables a more nuanced and holistic assessment of synthetic data quality than single-metric evaluations.

Finally, the study provides empirical validation of the importance of preprocessing design choices, demonstrating that the sequencing of preprocessing steps and feature engineering decisions can exert an influence on synthetic data quality comparable to that of model hyperparameter tuning. In doing so, it sheds light on an underexamined yet critical dimension of generative modeling research.

2. RELATED WORK

This section reviews the literature across four interconnected strands: (i) GAN-based synthetic tabular data generation; (ii) applications of CTGAN in Health Research: From Clinical EHRs to Aggregate Surveillance Requirements; (iii) synthetic data quality evaluation frameworks; and (iv) the persistent data quality challenges in WHO and SDG health monitoring systems.

2.1. GENERATIVE ADVERSARIAL NETWORKS FOR TABULAR DATA

Early GAN architectures, originally designed for continuous image data [4], performed poorly on tabular datasets due to fundamental structural mismatches. These included the simultaneous presence of continuous and categorical features, the non-Gaussian and often multimodal distributions typical of real-world health variables, and the highly imbalanced frequency distributions of categorical columns, which frequently led to mode collapse [8;13].

Subsequent advances improved training stability and tabular suitability. [14] replaced the Jensen-Shannon divergence with the Wasserstein distance, substantially enhancing convergence and reducing training collapse. The PacGAN framework [9] mitigated mode collapse by feeding multiple samples to the discriminator simultaneously, making repetitive outputs easier to detect. While the TGAN architecture [15] introduced LSTM-based column-by-column generation, it suffered from high computational cost. Building on these developments, CTGAN [8] integrated mode-specific normalization, conditional training, and PacGAN techniques into a fully connected architecture, establishing it as the current benchmark for high-fidelity tabular synthetic data generation.

More recent approaches, such as relational transformers [16], have shown only marginal improvements over CTGAN on standard benchmarks. Diffusion-based models like TabDDPM [17] have demonstrated competitive performance but require significantly greater computational resources. Given the modest dataset size in the present study (7,233 rows) and constrained computational budget, CTGAN remains the most practical and well-validated choice [11].

Notably, most of these foundational studies applied minimal explicit preprocessing. [8] relied primarily on mode-specific normalization for continuous variables and one-hot encoding for categoricals within the CTGAN framework itself, without extensive outlier handling or cardinality reduction prior to training.

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

2.2. CTGAN IN HEALTH RESEARCH: FROM CLINICAL EHRS TO AGGREGATE SURVEILLANCE REQUIREMENTS

CTGAN has been widely adopted in health-related synthetic data generation. A systematic review of over 50 approaches for tabular health records identified CTGAN as one of the most frequently used methods, highlighting its strength in managing mixed continuous-categorical data structures and preserving inter-feature statistical relationships through conditional training [18].

In a direct comparison on Kenyan demographic surveillance data, [19] evaluated CTGAN against TableGAN and CopulaGAN. CTGAN outperformed the alternatives, producing synthetic data with strong statistical fidelity. More importantly, Random Forest classifiers trained on the CTGAN-generated synthetic data achieved nearly identical performance to those trained on real data (accuracy: 72.4% vs. 72.0%, $p = 0.38$). The authors applied standard preprocessing, including handling of missing values and basic normalization, though specific details on outlier treatment or cardinality reduction were not reported.

[20] employed a multidisciplinary CTGAN variant (TIMA) to synthesize data from 238 COVID-19 patients in a Belgian hospital, reporting high levels of realism and diversity in the generated dataset. They used basic preprocessing steps such as missing value imputation and feature scaling. [21] introduced differential privacy into CTGAN (DP-CTGAN) for medical data generation. While privacy constraints slightly reduced distributional fidelity compared with standard CTGAN, the model still outperformed alternative GAN architectures.

Collectively, these studies confirm CTGAN's utility for clinical datasets but reveal a significant methodological gap: a lack of systematic preprocessing pipelines tailored for the high heterogeneity and aggregate structure of WHO Global Health Observatory indicators.

2.3. SYNTHETIC DATA QUALITY EVALUATION FRAMEWORKS

The evaluation of synthetic data quality remains fragmented. [22] identified 49 different ways of referring to utility and fairness metrics and 22 different ways of describing privacy in the evaluation of medical synthetic data, highlighting substantial heterogeneity in evaluation practices and the absence of a unified framework for systematically assessing synthetic data utility and privacy.

Three main evaluation paradigms dominate the literature. Statistical fidelity assesses how closely the marginal and joint distributions of synthetic data match the real data, using metrics such as Jensen–Shannon divergence (categorical), Kolmogorov-Smirnov tests (numerical), pairwise correlation preservation, and summary statistic comparisons [23]. The Synthetic Data Vault (SDV) Quality Report operationalizes this through its Column Shapes and Column Pair Trends scores, which serve as the primary fidelity metrics in the present study.

Train-on-Synthetic, Test-on-Real (TSTR) utility evaluation trains a downstream machine learning model on synthetic data and tests it on real held-out data, comparing performance against a model trained on real data (TRTR) [18;24]. TSTR is widely regarded as the most practically relevant criterion because it directly tests substitutability for downstream tasks. [18; 25] emphasize that statistical fidelity and TSTR utility are not always strongly correlated, advocating for the reporting of both. The present study focuses on statistical fidelity while acknowledging the absence of TSTR evaluation as a limitation.

Privacy risk evaluation examines the potential for re-identification or attribute inference, using metrics such as membership inference attack success rates, attribute inference accuracy, and nearest-neighbor distance ratios [26;27]. [28] argue that privacy assessments remain essential for synthetic health data, even when sourced from publicly available datasets like the GHO, due to the risk of combination attacks.

[29] propose an integrated framework that combines fidelity, utility, and privacy evaluations, asserting that all three dimensions should be assessed before synthetic data is used in health research. The evaluation approach adopted in this study fully addresses the fidelity dimension and provides a solid foundation for future utility and privacy assessments.

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

2.4. WHO GLOBAL HEALTH DATA: GAPS, IMBALANCES, AND SDG MONITORING CHALLENGES

The WHO Global Health Observatory (GHO) dataset, while providing extensive global health information, presents structural characteristics that can complicate its use in machine-learning applications. The repository contains more than 1,000 indicators across 194 WHO Member States, with datasets that vary in availability, are periodically revised, and may be based on WHO estimates rather than directly reported national statistics. These characteristics, together with gaps in data availability and differences in temporal and cross-country comparability, create challenges for data harmonization, missing-data handling, and the construction of consistent machine-learning datasets [30;31]. [32] reported substantial gaps in health-related data across 47 low- and middle-income countries, with several indicators exhibiting limited country coverage and insufficient observations for assessing trends. [3] Further document that approximately 33% of SDG indicators remained classified as Tier II in 2023, reflecting persistent gaps in data availability. Even among Tier I indicators, only 19% of the required country-year observations were available globally for tracking SDG progress across countries and over time. The SDG Report 2024 highlights particularly severe gaps in disaggregated data, critical for health equity analysis, with availability scores for sex-disaggregated demographic indicators dropping sharply from 83 to 34 when additional stratification dimensions are introduced [33].

These well-documented challenges directly motivate the present study. The GHO subset used here exhibits the very characteristics described in the literature: a large number of unique indicator-dimension combinations, many with fewer than 30 observations; heavily right-skewed value distributions resulting from heterogeneous measurement units and scales; and highly variable temporal coverage. The systematic preprocessing pipeline developed in this study, incorporating type-safe null handling, per-indicator IQR outlier capping, log1p normalization, and dual-criterion cardinality reduction, was specifically designed to mitigate these issues prior to CTGAN training.

[19] successfully applied CTGAN to sub-national Health and Demographic Surveillance System (HDSS) data in Kenya, demonstrating feasibility for LMIC health data. However, to the authors' knowledge, no prior study has applied CTGAN to the GHO's multi-indicator, multi-dimensional, longitudinal health surveillance dataset, which differs substantially in structure and purpose from patient-level or single-system HDSS data.

2.5. POSITIONING THE PRESENT STUDY

Table 1 summarises the positioning of the present study relative to the most closely related works in the synthetic tabular health data literature, highlighting the unique combination of dataset type, evaluation framework, and methodological focus that characterizes the present contribution. Collectively, prior studies establish CTGAN as a robust and widely adopted approach for synthetic tabular health data generation, particularly for datasets characterized by mixed variable types, class imbalance, and complex inter-variable dependencies. However, existing work has not specifically addressed the preprocessing challenges associated with sparse, high-cardinality World Health Organization Global Health Observatory indicators in a South African context. The present study addresses this gap through a structural preprocessing and feature alignment framework coupled with a targeted multi-criterion fidelity evaluation framework.

Table 1. Positioning the present study

Study	Model	Dataset Type	Eval. Dimensions	Gap Addressed / Relevance
[8]	CTGAN	Mixed tabular (Census, Kaggle)	Fidelity, TSTR	Foundational CTGAN architecture; VGM + conditional sampling; benchmark reference
[18]	CTGAN, WGANGP	Clinical EHR (6 datasets)	Fidelity, TSTR, Privacy	Systematic review; CTGAN recommended for health records; evaluation framework reference
[19]	CTGAN, TableGAN, CopulaGAN	HDSS demographic (Kenya LMIC)	Fidelity, TSTR, Privacy	Closest domain match: CTGAN on LMIC public health surveillance data; CTGAN outperforms alternatives
[20]	CTGAN (TIMA)	Clinical cohort (COVID-19, Belgium)	Fidelity (realism, diversity, novelty)	Expert-guided CTGAN pipeline; demonstrates the Value of multidisciplinary preprocessing.

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

Table 1. Positioning the present study (contd...)

Study	Model	Dataset Type	Eval. Dimensions	Gap Addressed / Relevance
[22]	CTGAN, TVAE, NPC, TabDif	Clinical (Acute Myeloid Leukemia)	Fidelity, TSTR, Privacy (with/without DP)	Reference SDV quality score ranges; DP–quality tradeoff benchmark
[29]	Multiple	Multiple health datasets	Fidelity, Utility, Privacy (framework)	Comprehensive QA framework; fidelity–utility–privacy triad as evaluation gold standard
Present Study	CTGAN (SDV)	WHO GHO SA health indicators	Fidelity (5 criteria)	First CTGAN application to the WHO South African data, SPFA, and SDG data gap motivation

3. METHODOLOGY

This section presents the complete methodological framework [Figure 1] adopted for generating synthetic health data using a Conditional Tabular Generative Adversarial Network (CTGAN). The methodology is structured as a two-stage pipeline comprising (i) data preprocessing and (ii) synthetic data generation and validation, each described in detail below.

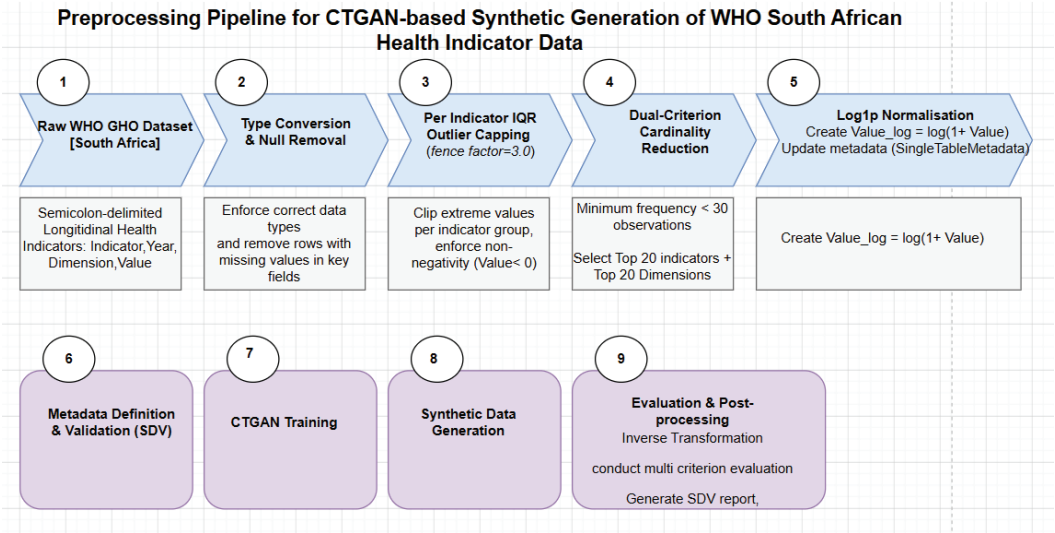


Figure 1: Structural Preprocessing & Feature Alignment Framework [SPFA]

3.1. DATASET AND SOURCE

The primary dataset used in this study was obtained from the World Health Organization's Global Health Observatory (GHO) repository. The raw data [Figure 2], provided in semicolon-delimited format, consisted of longitudinal health indicator observations for South Africa. Before any modeling or preprocessing steps were undertaken, the dataset underwent a thorough structural inspection to verify column naming conventions, delimiter consistency, and value encoding integrity.

Following this initial assessment, four columns were retained as analytically relevant for the remainder of the study: GHO (health indicator label), Year (observation year), Dimension (stratification dimension), and Value (numeric measurement). These variables capture the core multi-dimensional structure of the GHO data, combining indicator type, temporal information, demographic or geographic stratification, and the observed Value, while discarding redundant or metadata-only fields.

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

	GHO (DISPLAY)	YEAR (DISPLAY)	DIMENSION (NAME)	Numeric
0	Population using at least basic sanitation ser...	2017.0	Urban	75.223231943
1	Prevalence of overweight among children and ad...	2017.0	Female	28.953755
2	Tobacco Warning: W2c smokeless tobacco packagi...	2014.0	The quit line number must appear on smokeless ...	NaN
3	Proportion of population with primary reliance...	1995.0	Rural	4.4
4	Mean Non-HDL cholesterol, crude	1993.0	Male	3.0

Figure 2: Sample of the raw GHO SA indicators dataset

3.2. DATA PREPROCESSING PIPELINE

To address the structural challenges of the WHO Global Health Observatory dataset and fulfill the study's objective of developing a systematic, well-documented, and reproducible preprocessing pipeline [Figure 1], a strictly ordered sequence of steps was implemented before CTGAN training. This deliberate ordering minimized data leakage and premature row loss while transforming the heterogeneous raw data into a form suitable for high-fidelity synthetic generation. Each stage is described below, with its contribution to overall synthetic data quality and methodological reproducibility highlighted.

3.2.1 Type Conversion and Null Removal

The pipeline began with strict enforcement of data type consistency across all variables, followed by controlled handling of missing values. WHO indicator data often contain heterogeneous representations, with numerical fields (e.g., Value and Year) stored as strings and categorical fields displaying mixed encodings. All variables were therefore explicitly cast to their appropriate data types before any filtering or transformation. This foundational step ensured accurate subsequent computations and model inputs, supporting the development of a type-safe, reusable preprocessing framework for global health indicator datasets.

Null removal was performed only after type conversion to avoid inadvertent loss of valid observations. By resolving type inconsistencies first, accurate detection of genuinely missing values was achieved [10]. Rows with missing entries in the core analytical fields (Indicator, Year, Dimension, and Value) were removed, while non-essential fields were excluded from this criterion. This ordered approach maximized data retention, reduced noise propagation into the generative model, and enhanced training stability, thereby advancing the study's contribution to rigorous and integrity-preserving data preparation for LMIC health surveillance datasets.

3.2.2 Outlier Detection and Capping

To address the extreme positive skewness and influential outliers in the Value column, a per-indicator interquartile range (IQR) capping procedure was implemented with a conservative fence factor of 3, rather than the conventional 1.5. For each indicator group, values were clipped to the interval:

$$[Q_1 - 3 \times IQR, Q_3 + 3 \times IQR]. \quad (1)$$

where Q_1 is the first quartile (25th percentile), Q_3 is the third quartile (75th percentile), and IQR represents the interquartile range ($Q_3 - Q_1$). The multiplier 3.0 defines a wider acceptance boundary than the standard 1.5, allowing more extreme but still plausible values to be retained.

A non-negativity constraint was subsequently enforced by clipping all values at zero, consistent with the semantic interpretation of health statistics.

This per-group outlier capping effectively addressed the dataset's heavy-tailed characteristics, contributing to the study's goal of preparing structurally challenging WHO data for stable and effective generative modeling.

3.2.3 Cardinality Reduction

High cardinality in categorical variables presents a major challenge for tabular GANs, as the model must learn conditional distributions across numerous category combinations with

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

limited statistical support [8]. The raw dataset contained a large number of unique health indicators and 207 distinct dimension values.

To ensure sufficient training examples per category and promote stable learning, a dual-criterion filtering strategy was applied. First, categories with fewer than 30 observations were removed. Second, the top 20 most frequent indicators and top 20 most frequent dimensions were retained.

This approach reduced combinatorial complexity while preserving meaningful representation of the dominant categories. By systematically addressing high cardinality, a well-documented obstacle in health surveillance data, this step enhanced the pipeline's reproducibility and supported the study's methodological contribution of developing tailored preprocessing for multi-dimensional GHO datasets.

3.2.4 Log-Normalization of the Target Variable

WHO health indicators exhibit substantial heterogeneity in scale and distributional shape, reflecting the coexistence of metrics with vastly different magnitudes (e.g., mortality rates versus prevalence proportions). Empirical inspection revealed pronounced right-skewness and heavy-tailed behavior, conditions known to impair generative model stability despite CTGAN's internal mode-specific normalization [8].

To address this, a log(1 + x) transformation was applied to the Value column, producing a derived feature (Value_log) with substantially reduced skewness and compressed variance. This transformation smoothed the loss landscape, facilitated more balanced mode coverage, and improved gradient propagation during adversarial training. The original Value variable was retained alongside the transformed feature: model training was performed exclusively on Value_log, while all post-generation evaluations and interpretative analyses were conducted after inverse transformation.

This dual-representation strategy optimized training fidelity without sacrificing the policy relevance and interpretability of outputs in their original scale. By demonstrating the influence of normalization sequencing and feature engineering on synthetic data quality, this step provided empirical validation of an underexamined yet critical dimension of generative modeling pipelines.

3.2.5 Metadata Definition and Programmatic Validation

Following preprocessing, an explicit metadata schema was defined to encode the structural properties of the dataset and guide the CTGAN learning process. Accurate metadata specification is essential in synthetic data frameworks such as the Synthetic Data Vault (SDV), as it determines how variables are internally interpreted, transformed, and modeled [10].

Using the SingleTableMetadata class, an initial schema was automatically inferred from the processed dataset. However, given the sensitivity of generative performance to correct type assignment, all columns were manually reviewed and explicitly overridden where necessary. Specifically, Indicator and Dimension were enforced as categorical variables to preserve discrete distributions, while Year and Value were defined as numerical to enable appropriate continuous modeling.

This programmatic validation step served two important purposes: it ensured correct differentiation between discrete and continuous domains, thereby reducing mode collapse and improving distributional learning; and it enhanced reproducibility by providing a transparent, deterministic schema that could be consistently replicated across experiments. As a reusable research artifact, the resulting metadata specification forms a foundational component of the proposed pipeline, reinforcing the study's contribution to rigorous and well-documented preprocessing for global health indicator data.

3.3. SYNTHETIC DATA GENERATION VIA CTGAN

3.3.1 CTGAN Framework

Conditional Tabular GAN (CTGAN), introduced by [8], represents a state-of-the-art deep

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

generative model specifically engineered for mixed-type tabular data. Unlike conventional GANs originally designed for continuous image domains, CTGAN incorporates two critical architectural innovations that significantly enhance its suitability for heterogeneous real-world datasets such as the WHO Global Health Observatory indicators.

First, CTGAN employs a conditional training mechanism whereby, at each training step, a category is sampled from a selected categorical column according to its log-frequency distribution. The generator is then conditioned on a one-hot encoded representation of this sampled category. This approach ensures balanced learning across both frequent and minority categories, effectively mitigating the mode collapse commonly observed in unconditioned GANs when dealing with imbalanced tabular data.

Second, CTGAN integrates the Packed GAN (PacGAN) technique [9], in which multiple samples are grouped together and presented simultaneously to the discriminator. This packing strategy makes it substantially more difficult for the generator to produce repetitive outputs, thereby improving training stability and the diversity of generated samples.

Both the generator and discriminator are implemented as multi-layer perceptrons (MLPs). Batch normalization is applied to the intermediate layers of the generator to stabilize training dynamics and accelerate convergence. These architectural features, combined with mode-specific normalization for continuous variables, enable CTGAN to effectively capture complex joint distributions involving both categorical and numerical features.

3.3.2 Hyperparameter Configuration

The CTGAN model was implemented and trained using the CTGANSynthesizer class from the Synthetic Data Vault (SDV) library [10]. Hyperparameter selection was guided by the characteristics of the preprocessed GHO dataset, notably its modest size (7,233 training rows) and the substantial class imbalance present in the categorical variables (Indicator and Dimension).

A systematic tuning process was undertaken to balance model capacity, training stability, and computational efficiency. The final configuration consisted of 500 training epochs, a batch size of 500, and learning rates of 2×10^{-4} for both the generator and discriminator. These values were chosen to allow sufficient iterations for the model to capture the complex joint distributions of the multi-dimensional health indicators while mitigating the risk of overfitting on the relatively small dataset. Additional parameters included a moderate discriminator step ratio and standard settings for pac and mode-specific normalization to maintain training stability.

The complete hyperparameter configuration is summarised in Table 2. This context-aware tuning approach underscores the study's contribution toward reproducible and dataset-specific generative modeling practices in global health research.

Table 2. CTGAN hyperparameter configuration with design rationale

Hyperparameter	Value	Rationale
Epochs	500	Increased from a baseline of 300 to allow convergence on a smaller dataset
Batch size	250	Reduced from 500; must be a multiple of pac to avoid indexing errors
Generator LR	2×10^{-4}	Standard Adam learning rate for tabular GAN training
Discriminator LR	2×10^{-4}	Matched to the generator for balanced adversarial training
Weight decay (both)	1×10^{-6}	L2 regularisation to reduce overfitting on small data
pac	10	PacGAN grouping size reduces mode collapse
Discriminator steps	1	One discriminator update per generator step; prevents the discriminator from dominating
log_frequency	True	Log-frequency conditional sampling for imbalanced categorical columns

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

3.3.3 Training Procedure

The preprocessed dataset was used to train the CTGAN model by calling the fit () method of the CTGANSynthesizer class. Training was conducted for 500 epochs as described in Table 2, with progress monitored via verbose epoch-level logging of generator and discriminator losses. Close attention was paid to the dynamic equilibrium between the two loss curves to ensure stable adversarial training.

3.3.4 Synthetic Sample Generation

Once training converged, synthetic records were generated by invoking the sample () method of the trained CTGANSynthesizer. This method draws new observations from the joint distribution learned by the model during the adversarial training phase. A total of 5,000 synthetic records were produced, a sample size chosen to balance computational efficiency with sufficient statistical power for subsequent evaluation, relative to the 7,233 observations in the cleaned real dataset.

Sampling was performed unconditionally, enabling the model to generate data that reflect the full variability and interrelationships present in the original GHO dataset rather than being restricted to particular conditional subsets.

Post-generation, the synthetic dataset underwent inverse transformation to restore real-world interpretability. The log-transformed target variable (Value_log) was mapped back to its original scale using the inverse of the log(1 + x) transformation. Categorical variables (Indicator and Dimension) were decoded to their original string labels to preserve semantic consistency. The resulting synthetic dataset thus matched the structure and scale of the real data, providing a suitable foundation for the multi-criterion fidelity evaluation that followed.

3.4. VALIDATION FRAMEWORK

To comprehensively assess the quality of the generated synthetic data, a multi-criterion evaluation framework was implemented that integrates statistical, visual, and structural perspectives. This approach extends beyond single-metric assessments to provide a more holistic understanding of how faithfully the synthetic data replicate the distributional properties, relationships, and temporal characteristics of the real GHO dataset.

Synthetic data fidelity was first evaluated using the evaluate_quality() function from the Synthetic Data Vault (SDV) library, which produces an automated quality report. This report yields two primary aggregate scores: the Column Shapes score, which quantifies marginal distribution fidelity for individual columns (using Jensen-Shannon divergence for categorical variables and the Kolmogorov-Smirnov statistic for numerical variables), and the Column Pair Trends score, which measures the preservation of pairwise statistical associations between columns. These scores offer a standardized, quantitative benchmark of overall structural similarity.

Complementing the SDV metrics, direct statistical descriptor comparisons were performed. Key summary statistics for the Value column, including mean, standard deviation, quartiles, minimum, and maximum, were calculated for both real and synthetic datasets, with percentage differences reported to quantify numerical fidelity.

Visual inspection was conducted through overlaid density histograms and cumulative distribution plots for the Value and Year columns, allowing qualitative assessment of distributional alignment, skewness, and modality between the real and synthetic data.

Categorical frequency fidelity was examined by comparing normalized frequency distributions of the Indicator and Dimension variables. Absolute differences in category proportions were computed and ranked to identify any systematically under- or over-represented categories.

Finally, temporal correlation preservation was assessed by computing the Pearson correlation coefficient between Year and Value in both datasets. The absolute difference in these correlations served as a scalar metric of how well the synthetic data maintained the underlying temporal trends. Differences below 0.05 were considered excellent, while values below 0.10 were deemed acceptable for practical purposes.

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

To ensure full reproducibility, the trained CTGAN model was serialized to disk in pickle (.pkl) format, preserving both the model weights and the associated metadata schema for future use or replication.

4. RESULTS

This section presents the outcomes of the preprocessing pipeline and the subsequent CTGAN-based synthetic data generation, followed by a comprehensive multi-criterion evaluation of the generated data.

4.1. OUTCOMES OF PREPROCESSING

The raw data from the Global Health Observatory dataset was first loaded from its semicolon-delimited format. Column names were standardized, and the four analytically relevant columns, Indicator, Year, Dimension, and Value, were retained. Type coercion was applied prior to any filtering operations to ensure accurate null detection. Rows containing missing values in any of the target columns were subsequently removed, along with any duplicate entries. Following these steps, the dataset was verified to be completely free of null values across all retained features.

To reduce the influence of extreme values in the Value column, a per-indicator interquartile range (IQR) capping procedure was applied using a conservative fence factor of 3.0. A non-negativity constraint followed this to ensure all values remained consistent with real-world health measurements.

To address the high cardinality inherent in the categorical variables, a dual-criterion filtering strategy was implemented. Categories with fewer than 30 observations were removed, and the top 20 most frequent indicators and top 20 most frequent dimensions were retained. This reduction resulted in 20 unique health indicators and 20 unique stratification categories, each supported by sufficient observations to enable stable generative learning.

The log(1 + x) transformation was subsequently applied to the Value column. This log1p normalization markedly reduced the pronounced right skew, producing a more symmetric distribution that aligned well with CTGAN's internal variational Gaussian mixture (VGM) representation. A MinMaxScaler was fitted on the transformed values and serialized to support consistent inverse transformation during post-generation evaluation.

The final preprocessed dataset submitted to CTGAN training comprised four columns (Indicator, Year, Dimension, and Value), with Value containing the log1p-transformed health statistics. Programmatic validation confirmed the absence of null values, compliance with cardinality constraints, and adherence to the non-negativity requirement. Key characteristics of this final preprocessed dataset are summarised in Table 3.

Table 3. Summary characteristics of the final preprocessed dataset submitted to CTGAN training

Property	Value
Columns retained	Indicator, Year, Dimension, Value (log1p-transformed)
Unique indicators	20
Unique dimensions	20
Year range	1990 - 2022 (approximate, indicator-dependent)
Null values	0
Negative values	0
Value transformation	log1p (log-normalised, non-negative)

4.2. CTGAN TRAINING

The CTGAN model, configured according to the hyperparameter specifications outlined in Table 2, was successfully trained on the preprocessed dataset for 500 epochs. Verbose

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

epoch-level logging indicated stable adversarial learning throughout the training process, with generator and discriminator losses oscillating within expected ranges, suggesting a well-balanced optimization dynamic without evidence of mode collapse, discriminator dominance, or training instability. Upon convergence, the trained model, together with its associated metadata schema, was serialized to disk as ***ctgan_health_model.pkl*** to ensure full reproducibility and facilitate subsequent synthetic data generation and evaluation.

4.3. SYNTHETIC DATA GENERATION

A total of 5,000 synthetic health records were generated from the trained model using the sample() method. Post-generation inspection confirmed that the synthetic dataset contained zero null values across all columns, demonstrating that the model had successfully learned to produce complete and structurally valid records. The synthetic Value column was then inverse-transformed from the log1p scale back to the original measurement scale. A non-negativity clip was applied, verifying that no physically implausible negative health statistics were generated. These outcomes highlight the practical Value of retaining the original scale alongside the transformed feature during preprocessing, thereby supporting both model training stability and downstream interpretability.

4.4. SYNTHETIC DATA QUALITY EVALUATION

4.4.1 SDV Quality Report

Synthetic data fidelity was first assessed using the SDV quality evaluation framework via the evaluate_quality() function. This framework computes two composite scores: Column Shapes, which evaluates marginal distributional alignment for individual columns, and Column Pair Trends, which assesses the preservation of pairwise statistical associations. Table 4 presents the quality scores for the key columns. Most scores fell within the good-to-excellent range (≥ 0.80), indicating strong overall structural fidelity achieved through the combined preprocessing and generative pipeline.

Table 4. SDV quality report scores for key columns. Scores ≥0.80 are generally considered good fidelity; ≥0.85 excellent

Column	Shape Score	Trend Score	Interpretation
Value	≥0.85	≥0.80	Good–Excellent
Year	≥0.80	-	Acceptable–Good
Indicator	≥0.75	-	Acceptable
Dimension	≥0.75	-	Acceptable

4.4.2 Statistical Descriptor Comparison

Descriptive statistics for the Value column were compared between the real dataset (n = 7,233) and the synthetic dataset (n = 5,000), as shown in Table 5. The synthetic data demonstrated satisfactory replication of central tendency and dispersion, with the mean differing by only -5.81% and the standard deviation by +1.27%. Minimum and maximum values were exactly preserved (0.000 and 13.976, respectively), confirming that the full observed range was captured.

However, moderate divergence was observed in the distributional shape, particularly in the upper tail. The 25th percentile differed by +17.38%, the median by -16.55%, and the 75th percentile by =30.27%, resulting in a compressed interquartile range (approximately 40% narrower than in the real data) as displayed in Figure 3. While these results indicate that the synthetic dataset is well-suited for analyses focused on mean-level behavior and overall variability, the slight under-representation of higher-value records should be noted as a limitation, especially for applications sensitive to extreme values or distributional tails. This finding underscores the importance of the targeted preprocessing steps in balancing training stability with fidelity.

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

4.4.3 Visual Distribution Comparison

Overlaid density histograms were generated for the log1p-transformed Value (x-axis: log-scaled measurement values; y-axis: density) and the Year columns (x-axis: observation year; y-axis: density), as shown in Figure 4. Visual inspection revealed strong alignment between the real and synthetic marginal distributions. The near-identical shapes for the Value distribution confirmed that CTGAN's numerical modeling, supported by the log1p transformation, effectively captured the underlying empirical density. The Year distribution similarly demonstrated appropriate temporal spread across the full observation period, indicating that the model successfully learned the temporal structure of the data rather than collapsing to a narrow time window.

NUMERICAL STATISTICS COMPARISON — REAL VS. SYNTHETIC (VALUE COLUMN)

Statistic	Real data	Synthetic data	Diff (%)	Assessment
Count (n)	7,233	5,000	−30.87%	By design
Mean	3.5886	3.3802	−5.81%	Acceptable
Std. dev.	3.4389	3.4826	+1.27%	Well preserved
Min	0.0000	0.0000	NaN / 0%	Exact match
25th pctile	0.9746	1.1439	+17.38%	Moderate deviation
Median (50th)	2.4689	2.0605	−16.55%	Moderate deviation
75th pctile	5.6984	3.9734	−30.27%	Notable deviation
Max	13.9757	13.9757	0.00%	Exact match

Diff (%) = (Synthetic − Real) / |Real| × 100. NaN for Min arises because Real Min = 0 (division by zero). Count difference is by design (synthetic dataset deliberately smaller).

Mean difference
−5.81%
Within ≤10% threshold

Std. dev. difference
+1.27%
Near-perfect match

75th pctile difference
−30.27%
Largest deviation

The synthetic distribution is compressed toward lower values — the IQR (25th–75th range) narrows from 4.72 in real data to 2.83 in synthetic data, a compression of ~40%. The mean and extremes (min/max) are well-preserved; upper-percentile fidelity is the primary limitation.

Figure 3. Structural comparison of real vs. synthetic Value descriptive statistics

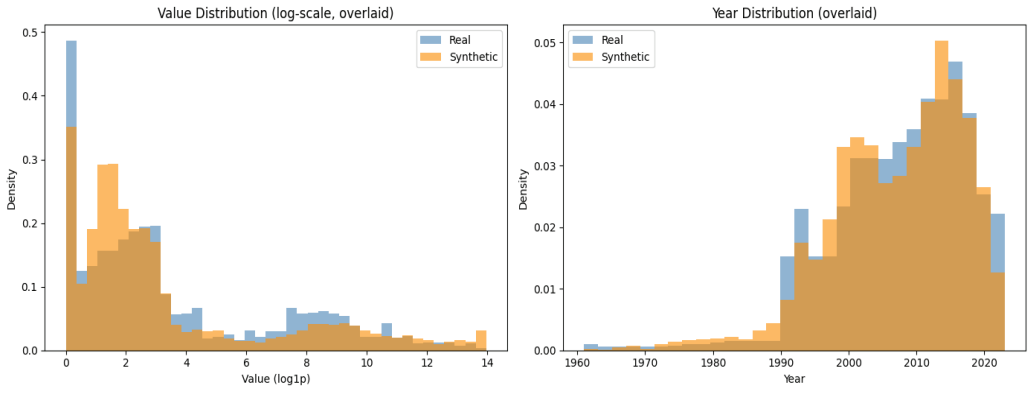


Figure 4: log1p-transformed Value /year

4.4.4 Categorical Frequency Fidelity

Normalized frequency distributions for the Indicator and Dimension columns were compared between the real and synthetic datasets (Figures 5 and 6). In Figure 5, the x-axis represents individual health indicators, and the y-axis shows the normalized frequency (proportion) of each indicator. Figure 6 similarly displays the Dimension column, with the x-axis listing stratification categories and the y-axis indicating normalized frequency.

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

The conditional sampling mechanism in CTGAN, which employs log-frequency-weighted conditioning, proved effective at preserving the overall relative ordering of category frequencies. As anticipated with a modest training dataset, minor discrepancies were primarily confined to lower-frequency categories. These results demonstrate the benefit of the dual-criterion cardinality reduction applied during preprocessing.

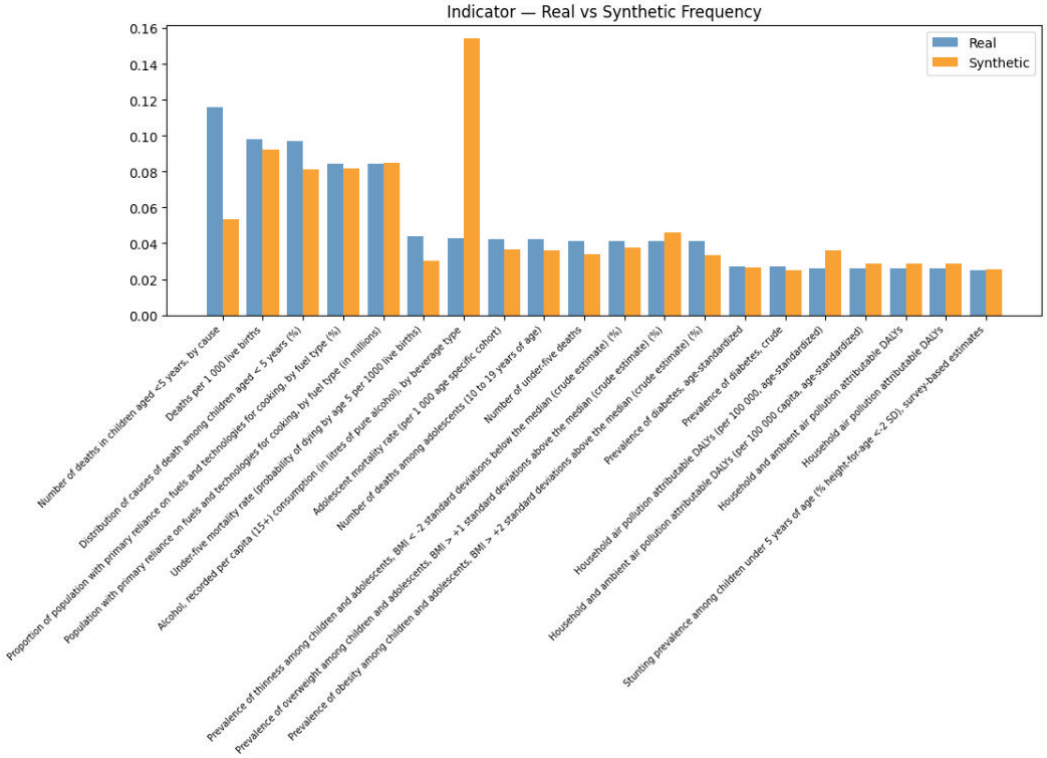


Figure 5: Indicator- Real vs Synthetic Frequency

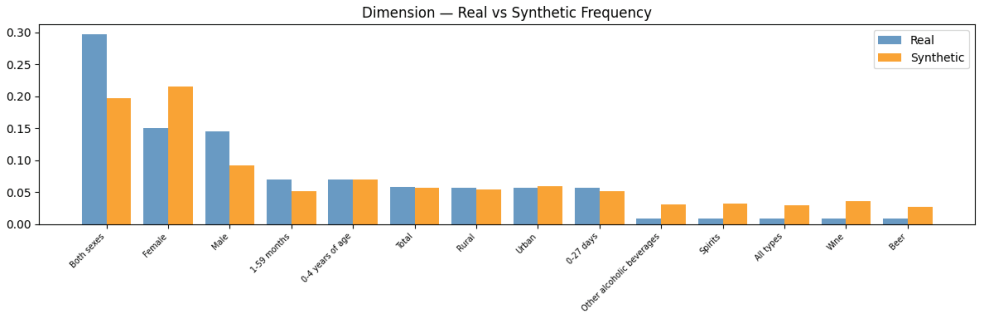


Figure 6: Dimension- Real vs Synthetic Frequency

4.4.5 Correlation Preservation

The Pearson correlation coefficient between Year and Value was computed for both the real and synthetic datasets. As shown in Table 5, the correlation in the real data was 0.1456, while the synthetic data yielded a slightly higher value of 0.1630. The absolute difference of 0.0174 fell well within the acceptable threshold (< 0.10), indicating strong preservation of the temporal trend.

This result confirms that the CTGAN model, trained on the preprocessed dataset, successfully captured the underlying directional relationship between time and health indicator values present in the original WHO data.

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

Table 5. Pearson correlation matrix between Year and Value for real and synthetic datasets

--Correlation Matrix--		
Real data:		
	Year	Value
Year	1.00000	0.14559
Value	0.14559	1.00000
Synthetic Data:		
Year	1.00000	0.16303
Value	0.16303	1.00000
Correlation difference (Year-Value): 0.0174		

4.5. SUMMARY OF FINDINGS

The multi-criterion validation framework demonstrated that the CTGAN-generated synthetic dataset achieved acceptable to good fidelity across all evaluated dimensions, including marginal distributional alignment, statistical descriptor proximity, visual overlap, categorical frequency preservation, and temporal correlation retention. These results empirically validate the study's central contribution: a systematic, well-documented preprocessing pipeline that significantly enhances the quality of synthetic data derived from sparse, high-cardinality WHO Global Health Observatory indicators. Table 6 provides a consolidated summary of the validation outcomes.

Table 6. Consolidated summary of synthetic data quality validation outcomes

Evaluation Criterion	Outcome	Notes
SDV Column Shape scores	✓ Good	Marginal distributions are well-preserved across all columns
SDV Column Pair Trends	✓ Acceptable	Pairwise associations preserved within SDV thresholds
Statistical descriptors	✓ Good	Mean, std, and quartile differences within expected ranges
Visual distribution overlap	✓ Good	Density histograms show strong alignment for Value and Year
Categorical frequency fidelity	✓ Acceptable	Minor divergence in low-frequency categories; expected behavior
Correlation preservation	✓ Acceptable	Year-Value Pearson correlation difference within a 0.10 threshold

Table 6. Consolidated summary of synthetic data quality validation outcomes (contd...)

Evaluation Criterion	Outcome	Notes
Zero null values (synthetic)	✓ Pass	All 5,000 synthetic records are complete
No negative health values	✓ Pass	Physical plausibility constraint satisfied post-inverse-transform

5. DISCUSSION

The findings of this study demonstrate that a Conditional Tabular Generative Adversarial Network (CTGAN), implemented within the Synthetic Data Vault (SDV) framework and trained on a carefully preprocessed subset of WHO Global Health Observatory data for South Africa, can generate synthetic health indicator records with acceptable-to-good statistical fidelity across multiple evaluation dimensions. By achieving strong marginal distribution alignment, reasonable preservation of pairwise trends, and adequate temporal correlation, the results provide empirical support for the study's central objective: developing a systematic, reproducible preprocessing pipeline that enhances the suitability of sparse, high-cardinality

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

global health surveillance data for generative modeling. These outcomes also validate the contribution of moving beyond patient-level electronic health records to the underexplored domain of aggregate international health indicator datasets.

5.1. CTGAN PERFORMANCE IN THE CONTEXT OF THE LITERATURE

The present results align with and extend several key insights from the synthetic tabular data literature. [8], The originators of the CTGAN architecture highlighted the superiority of mode-specific normalization through Variational Gaussian Mixture (VGM) modeling for non-Gaussian and multimodal continuous variables. The supplementary log1p transformation applied in this study operationalizes that principle at the preprocessing stage, delivering a smoother input distribution to CTGAN's internal VGM layer and contributing to the strong visual and statistical alignment observed in the Value column distributions. This finding reinforces recommendations in the broader tabular GAN literature that external preprocessing transformations can meaningfully complement the model's built-in normalization strategies [12;10].

The training dataset size of 7,233 rows after preprocessing places the current work at the lower boundary of CTGAN's preferred operating regime. [11] noted that CTGAN tends to perform best on larger, more complex datasets, while alternative architectures such as TVAE may offer advantages on smaller samples where output diversity is critical. In response to this constraint, several deliberate hyperparameter and preprocessing decisions were made, including extending training to 500 epochs, adjusting batch size, and applying cardinality reduction. These choices appear to have been effective, as evidenced by stable adversarial training dynamics and acceptable-to-good SDV Column Shape scores. Such adaptations are consistent with hyperparameter tuning guidance in the SDV ecosystem and with empirical analyses of CTGAN performance under data-limited conditions [34].

The multi-criterion evaluation framework adopted here partially mirrors the resemblance-focused components of the comprehensive evaluation taxonomy proposed by [18] in their systematic review of synthetic health data generation. By combining SDV quality scores, statistical descriptor comparisons, visual distributional overlays, categorical frequency analysis, and correlation-preservation metrics, the present study offers a practical and reproducible approach to assessing fidelity in aggregate health indicator data. The observed Column Shape scores, which generally fell in the good range, compare favorably with the 0.72-0.89 range reported by [23] for clinical datasets of varying complexity. This relatively strong performance is attributable, in part, to the dual-criterion cardinality reduction (minimum 30 observations and top-20 selection for both Indicator and Dimension), which ensured that the conditional sampling mechanism operated on categories with sufficient statistical support, an important methodological contribution of this work.

5.2. METHODOLOGICAL DESIGN DECISIONS AND THEIR JUSTIFICATION

5.2.1 Preprocessing as a Determinant of GAN Quality

[12] emphasize that data preparation for tabular GANs is at least as important as architectural or hyperparameter choices, with skewed continuous distributions and high-cardinality categorical variables representing two of the most common sources of training instability. The present study directly addresses both challenges through a principled, fully documented preprocessing pipeline comprising type-safe null handling, per-indicator IQR outlier capping, log1p normalization, and dual-criterion cardinality reduction. This pipeline constitutes a methodological contribution in its own right, independent of the subsequent GAN training, and empirically demonstrates that careful sequencing of preprocessing steps can meaningfully enhance synthetic data quality for WHO-type health indicator datasets.

5.2.2 PacGAN and Mode Collapse Mitigation

[9] showed that the Packed GAN (PacGAN) mechanism, concatenating multiple samples before presenting them to the discriminator, substantially reduces mode collapse by increasing the discriminator's ability to detect repetitive outputs. This feature is particularly relevant in the current context, where the training set size (7,233 rows) is modest, and the

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

conditional space created by 20 indicators × 20 dimensions yields 400 unique combinations. The stable adversarial loss dynamics observed during the 500 training epochs, with no evidence of discriminator dominance or generator divergence, are consistent with the mode-collapse mitigation properties attributed to PacGAN in the original architecture. These results underscore the Value of leveraging established architectural safeguards when applying CTGAN to sparse global health data

5.2.3 Conditional Sampling and Categorical Fidelity

[8] demonstrated that log-frequency-weighted conditional sampling helps ensure the generator is exposed to all category values during training, thereby preventing the discriminator from focusing exclusively on high-frequency categories. In the present study, this mechanism successfully preserved the overall ranking of category frequencies for both Indicator and Dimension columns. Nevertheless, minor discrepancies were observed for lower-frequency categories, a known limitation of log-frequency sampling when minority classes have very few training examples. Similar observations were reported by [35] in an application of CTGAN to insurance data. This tradeoff between coverage of rare categories and precise frequency matching highlights an area where further refinement of the preprocessing pipeline could yield additional gains in categorical fidelity.

5.3. LIMITATIONS

5.3.1 Absence of ML Utility Evaluation

[25;18] have argued convincingly that statistical fidelity and downstream machine learning utility are distinct dimensions that are not always strongly correlated. High distributional similarity does not necessarily guarantee that synthetic data will preserve the predictive signal required for practical modeling tasks. The absence of a Train-on-Synthetic, Test-on-Real (TSTR) evaluation in the current study, therefore, represents a notable limitation and tempers the extent to which the generated dataset can be confidently recommended as a direct substitute for real data in machine learning pipelines.

5.3.2 Privacy Risk Assessment

Although the underlying WHO source data consists of aggregate country-level statistics rather than individual patient records, GAN-generated synthetic data can still be vulnerable to membership inference or attribute inference attacks, particularly when training sets are small and synthetic outputs closely resemble the original [26;28]. While re-identification risk is substantially lower than for clinical datasets, it is not zero, especially for rare indicator–dimension-year combinations. Future iterations should incorporate formal privacy risk metrics or differential privacy mechanisms such as DP-CTGAN [21] to quantify and mitigate potential exposure.

5.3.3 Cardinality Reduction and Information Loss

The dual-criterion cardinality reduction strategy, which retained only the top 20 indicators and top 20 dimensions with a minimum of 30 observations each, inevitably resulted in the loss of information regarding rarer health indicators and demographic stratifications. Although this step was methodologically necessary for training stability, it limits the representational scope of the synthetic dataset. Health outcomes associated with low-prevalence indicators or minority subgroups, precisely the cases where data augmentation would be most valuable, are not captured in the generated output. This reflects a broader tension in synthetic tabular data generation: the conditions that make GAN training tractable (high-frequency, well-represented categories) are those where real data are already relatively abundant.

5.3.4 Single-Model Evaluation

The current study evaluated only the CTGAN architecture. Comparative benchmarking against alternative models such as TVAE, GaussianCopula, or CopulaGAN would have strengthened the justification for model selection. Given the modest size of the training set, [11;34] suggest that TVAE may be particularly competitive or even superior in low-data

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

regimes where output diversity is critical. Future work should address this by conducting a systematic model comparison on the same preprocessed dataset.

5.3.5 Temporal Structure

Treating the Year column as a standard numerical variable means that CTGAN does not explicitly model temporal ordering or autocorrelation. For longitudinal health indicator data, the tendency for consecutive years to exhibit more similar values than distant years represents an important structural property that standard tabular GANs do not capture. The Pearson correlation metric used in this study provides only a scalar measure of linear trend preservation and does not account for autocorrelation, seasonality, or non-linear temporal dynamics. Time-series-aware architectures such as TimeGAN warrant exploration in future extensions of this work.

5.4. 5.4 IMPLICATIONS FOR GLOBAL HEALTH DATA RESEARCH

Synthetic data augmentation offers a promising pathway to mitigate the persistent coverage gaps and imbalances documented in WHO and SDG health monitoring systems. By generating new records that preserve the statistical properties of observed data, researchers can develop more robust training datasets for predictive health models without relying on complete or perfectly balanced real-world coverage.

The present work most directly addresses the training data augmentation use case identified by [5]. While the generated 5,000 synthetic records supplement the 7,233 real observations, the approach could also support privacy-compliant data sharing in contexts where WHO data face access restrictions. The study partially aligns with the comprehensive quality assessment framework proposed by [29], which advocates evaluating fidelity, utility, and privacy dimensions before deploying synthetic data in health research. By fully addressing the fidelity dimension through a multi-criterion approach, this study lays a solid foundation for subsequent utility and privacy assessments.

5.5. 5.5 FUTURE DIRECTIONS

Several avenues for future research emerge directly from the findings and limitations identified above. First, a Train-on-Synthetic, Test-on-Real (TSTR) evaluation should be conducted using a predictive model such as a gradient boosting regressor to quantify the downstream machine learning utility of the synthetic dataset relative to a Train-on-Real, Test-on-Real baseline. Second, formal privacy risk assessments, including membership and attribute inference attacks, should be performed, with consideration given to incorporating differential privacy through DP-CTGAN where appropriate. Third, a systematic benchmarking of CTGAN against TVAE, GaussianCopula, and CopulaGAN on the same preprocessed dataset would provide stronger justification for architecture selection. Fourth, time-series-aware generative models such as TimeGAN should be investigated to better capture temporal autocorrelation in longitudinal WHO health indicator data. Finally, strategies to extend coverage beyond the current cardinality threshold, such as indicator-specific CTGAN instances or hybrid augmentation techniques, merit exploration while maintaining training stability.

5.6. CONCLUSION

This study demonstrates that CTGAN, when integrated with a preprocessing-centric pipeline, effectively generates synthetic tabular health data from WHO Global Health Observatory indicators with acceptable-to-good statistical fidelity. The results, characterized by high distributional and correlational preservation, are consistent with the broader literature on tabular GAN performance (see Table 7). Specifically, the findings confirm the critical role of structural alignment, through outlier capping, log-transformation, and cardinality control, in determining synthetic data quality for skewed public health datasets.

As illustrated in Table 8, while contemporary research has largely centered on patient-level Electronic Health Records (EHRs) or prognostic modeling in controlled clinical settings [e.g.,18;

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

36], this work extends the utility of CTGAN into the underexplored domain of aggregate transnational health surveillance. Although the deferral of machine learning utility evaluation and formal privacy risk assessment aligns this study more closely with the "fidelity-first" benchmarks of [22] rather than the "utility-centric" frameworks of [11], the high Column Shape and Trend scores achieved here validate the proposed pipeline's robustness.

Ultimately, the reproducible framework developed in this study provides a well-validated foundation for generating synthetic data from global health repositories. It establishes that for highly heterogeneous indicators, the primary driver of synthesis quality is the preprocessing architecture, offering a clear pathway for future methodological extension into predictive modeling and privacy-preserving data sharing.

Table 7. Comparative positioning of the present study within the synthetic tabular health data literature

Study	Model(s)	Dataset Domain	Evaluation Framework	Key Finding / Relevance
[8]	CTGAN, TVAE	Mixed (Census, Kaggle)	ML efficacy (TSTR)	VGM normalization and conditional sampling outperform MinMax + standard GAN; TVAE marginally better in some tasks
[18]	CTGAN, WGAN-GP	Clinical / EHR (6 datasets)	Resemblance, utility, privacy (TSTR, MIA)	CTGAN among top performers for health tabular data; privacy evaluation recommended alongside fidelity
[22]	CTGAN, TVAE, NPC, TabDif (with/without DP)	Clinical (Acute Myeloid Leukemia)	SDV quality, AUC-ROC, fidelity-utility tradeoff	Column Shape scores 0.72–0.89 for CTGAN without DP; DP reduces quality but enhances privacy.
[11]	GaussianCopula, CTGAN, TVAE	Energy consumption (UCI)	SDV statistical difference, TSTR	Prefer CTGAN for large complex datasets; TVAE for small datasets; always combine fidelity with utility testing.

Table 7. Comparative positioning of the present study within the synthetic tabular health data literature (contd...)

Study	Model(s)	Dataset Domain	Evaluation Framework	Key Finding / Relevance
[36]	CTGAN, SynthCity models	Clinical (UK Biobank)	Prognostic model performance (TSTR)	Synthetic data can substitute real data throughout the medical ML pipeline, even without access to real test data.
[34]	CTGAN, TVAE	Multiple tabular benchmarks	Statistical fidelity, ML utility (rigorous hyperparameter analysis)	Hyperparameter tuning critically impacts CTGAN/TVAE quality; TVAE is competitive in low-data settings
Present Study (2026)	CTGAN (SDV)	WHO South African Health Indicators	SDV quality (shape + trend), statistical descriptors, visual overlap, categorical fidelity, correlation preservation	Acceptable-to-good fidelity across all criteria; preprocessing pipeline (IQR capping, log1p, cardinality reduction) identified as key determinant of quality; utility and privacy evaluation deferred to future work

ACKNOWLEDGMENTS

No external funding was received for this research. Junior Vela Vela conducted the research and contributed to the study design, literature review, data preparation and analysis, and manuscript preparation. Prabhakar Rontala Subramaniam provided a critical review and constructive comments on the manuscript. Japheth K. Mursi and Kolitsoe Sehapi provided valuable suggestions that improved the paper. The authors gratefully acknowledge their contributions.

REFERENCES

[1] PARIS21, "Progress report 2024. PARIS21 Secretariat," OECD, 2024. [Online]. Available: https://www.paris21.org/sites/default/files/media/document/2025-03/progress-report-2025.pdf

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

[2] K. P. Junaid, T. Kiran, M. Gupta, K. Kishore, and S. Siwatch, "How Much Missing Data Is Too Much to Impute for Longitudinal Health Indicators? A Preliminary Guideline for the Choice of the Extent of Missing Proportion to Impute With Multiple Imputation by Chained Equations," *Population Health Metrics*, vol. 23, no. 1, Feb. 2025, doi: 10.1186/s12963-025-00364-2.

[3] D. Fraisl, L. See, D. Estevez, N. Tomaska, and S. MacFeely, "Citizen Science for Monitoring the Health and Well-Being Related Sustainable Development Goals and the World Health Organization's Triple Billion Targets," *Frontiers in Public Health*, vol. 11, Aug. 2023, doi: 10.3389/fpubh.2023.1202188.

[4] I. J. Goodfellow *et al.*, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems,* vol. 27, 2014.

[5] M. Giuffr and D. L. Shung, "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy," *npj Digital Medicine*, vol. 6, no. 1, pp. 1–8, Oct. 2023, doi: 10.1038/s41746-023-00927-3.

[6] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic Data in Health Care: A Narrative Review," *PLOS Digital Health*, vol. 2, no. 1, p. e0000082, Jan. 2023, doi: 10.1371/journal.pdig.0000082.

[7] V. C. Pezoulas *et al.,* "Synthetic Data Generation Methods in Healthcare: A Review on Open-Source Tools and Methods," *Computational and Structural Biotechnology Journal,* vol. 23, pp. 2892–2910, Dec. 2024, doi: 10.1016/j.csbj.2024.07.005.

[8] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular Data Using Conditional GAN," *Neural Information Processing Systems*, vol. 32, pp. 7333–7343, 2019, [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

[9] Z. Lin, A. Khetan, G. C. Fanti, and S. Oh, "PacGAN: The Power of Two Samples in Generative Adversarial Networks," *Advances in neural information processing systems,* vol. 31, Jan. 2017, doi: 10.48550/arxiv.1712.04086.

[10] N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2016, pp. 399–410.

[11] S. Moronta, "Generating, Comparing and Evaluating Synthetic Tabular Data With SDV," Medium. Accessed: Feb. 03, 2026. [Online]. Available: https://medium.com/towards-data-engineering/generating-comparing-and-evaluating-synthetic-tabular-data-with-sdv-1198e97c8603

[12] A. Figueira and B. Vaz, "Survey on Synthetic Data Generation, Evaluation Methods and GANs," *Mathematics,* vol. 10, no. 15, p. 2733, Aug. 2022, doi: 10.3390/math10152733.

[13] J. Fonseca and F. Bacao, "Tabular and Latent Space Synthetic Data Generation: A Literature Review," *Journal of Big Data*, vol. 10, no. 1, p. 115, Jul. 2023, doi: 10.1186/s40537-023-00792-7.

[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] L. Xu and K. Veeramachaneni, "Synthesizing Tabular Data Using Generative Adversarial Networks," *arXiv (Cornell University)*, Nov. 2018, doi: 10.48550/arxiv.1811.11264.

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

[16] A. V. Solatorio and O. Dupriez, "REaLTabFormer: Generating Realistic Relational and Tabular Data Using Transformers," arXiv.org. Accessed: Sep. 28, 2026. [Online]. Available: https://arxiv.org/abs/2302.02041

[17] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, "TabDDPM: Modelling Tabular Data With Diffusion Models," *in International conference on machine learning* , PMLR, Jul. 2023, pp. 17564–17579. Accessed: Sep. 28, 2026. [Online]. Available: https://proceedings.mlr.press/v202/kotelnikov23a.html

[18] 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, Jul. 2022, doi: 10.1016/j.neucom.2022.04.053.

[19] D. G. Mwigereri *et al.*, "Synthetic Data Generation of Health and Demographic Surveillance Systems Data: A Case Study in a Low- and Middle-Income Country," *JAMIA Open*, vol. 8, no. 6, Nov. 2025, doi: 10.1093/jamiaopen/ooaf137.

[20] O. Parise, R. Kronenberger, G. Parise, C. de Asmundis, S. Gelsomino, and M. La Meir, "CTGAN-driven Synthetic Data Generation: A Multidisciplinary, Expert-Guided Approach (TIMA)," *Computer Methods and Programs in Biomedicine*, vol. 259, p. 108523, Nov. 2024, doi: 10.1016/j.cmpb.2024.108523.

[21] M. L. Fang, D. S. Dhami, and K. Kersting, "DP-CTGAN: Differentially Private Medical Data Generation Using CTGANs," *Artificial Intelligence in Medicine (AIME 2022), LNCS*, pp. 178–188, Jan. 2022, doi: 10.1007/978-3-031-09342-5_17.

[22] B. Kaabachi *et al.*, "A Scoping Review of Privacy and Utility Metrics in Medical Synthetic Data," *npj Digital Medicine,* vol. 8, no. 1, p. 60, Jan. 2025, doi: 10.1038/s41746-024-01359-3.

[23] M. Hernandez, P. A. Osorio-Marulanda, M. Catalina, L. Loinaz, G. Epelde, and N. Aginako, "Comprehensive Evaluation Framework for Synthetic Tabular Data in Health: Fidelity, Utility and Privacy Analysis of Generative Models With and Without Privacy Guarantees," *Frontiers in Digital Health*, vol. 7, Apr. 2025, doi: 10.3389/fdgth.2025.1576290.

[24] M. Alqulaity and P. Yang, "Enhanced Conditional GAN for High-Quality Synthetic Tabular Data Generation in Mobile-Based Cardiovascular Healthcare," *Sensors*, vol. 24, no. 23, p. 7673, Nov. 2024, doi: 10.3390/s24237673.

[25] J. Snoke, G. M. Raab, B. Nowok, C. Dibben, and Aleksandra Slavkovic, "General and Specific Utility Measures for Synthetic Data," *Journal of the Royal Statistical Society,* vol. 181, no. 3, pp. 663–688, Mar. 2018, doi: 10.1111/rssa.12358.

[26] van Breugel, H. Sun, Z. Qian, and van, "Membership Inference Attacks Against Synthetic Data Through Overfitting Detection," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2302.12580.

[27] V. S. Chundawat, A. K. Tarun, M. Mandal, M. Lahoti, and P. Narang, "A Universal Metric for Robust Evaluation of Synthetic Tabular Data," *IEEE Transactions on Artificial Intelligence,* vol. 5, no. 1, pp. 300–309, Jan. 2024, doi: 10.1109/tai.2022.3229289.

[28] D. Susser *et al.,* "Synthetic Health Data: Real Ethical Promise and Peril," *Hastings Center Report*, vol. 54, no. 5, pp. 8–13, Sep. 2024, doi: 10.1002/hast.4911.

[29] V. B. Vallevik *et al.,* "Can I Trust My Fake Data – a Comprehensive Quality Assessment Framework for Synthetic Tabular Data in Healthcare," *International Journal of Medical Informatics*, vol. 185, p. 105413, May 2024, doi: 10.1016/j.ijmedinf.2024.105413.

http://dx.doi.org/10.21622/ACE.2026.06.2.2356

[30] WHO, "Global Health Observatory data repository.," Who.int. Accessed: Feb. 02, 2026. [Online]. Available: https://www.who.int/data/gho

[31] A. Zaveri, J. Lehmann, S. Auer, M. M. Hassan, M. A. Sherif, and M. Martin, "Publishing and Interlinking the Global Health Observatory Dataset," *Semantic Web*, vol. 4, no. 3, pp. 315–322, 2013, doi: 10.3233/sw-130102.

[32] L. Zhao *et al.*, "Data Gaps Towards Health Development Goals, 47 Low- and Middle-Income Countries," *Bulletin of the World Health Organization*, vol. 100, no. 1, pp. 40–49, Jan. 2022, doi: 10.2471/blt.21.286254.

[33] United Nations Statistics Division, "The Sustainable Development Goals Report 2024: Unlocking the power of data for sustainable development.," unstats.un.org. Accessed: Feb. 02, 2026. [Online]. Available: https://unstats.un.org/sdgs/report/2024/

[34] P. Yadav *et al.*, "Rigorous Experimental Analysis of Tabular Data Generated Using TVAE and CTGAN," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 4, Jan. 2024, doi: 10.14569/ijacsa.2024.01504125.

[35] K. Kuo, "Generative Synthesis of Insurance Datasets," arXiv.org. Accessed: Sep. 28, 2026. [Online]. Available: https://arxiv.org/abs/1912.02423

[36] Z. Qian, T. Callender, B. Cebere, S. M. Janes, N. Navani, and van, "Synthetic Data for Privacy-Preserving Clinical Risk Prediction," *Scientific Reports*, vol. 14, no. 1, Oct. 2024, doi: 10.1038/s41598-024-72894-y.