information

MDPI

Article

Generative Data Modelling for Diverse Populations in Africa: Insights from South Africa

Sally Sonia Simmons [1,*], John Elvis Hagan, Jr. [2,3,*] and Thomas Schack [2]

[1] Department of Social Policy, London School of Economics and Political Science, London WC2A 2AE, UK
[2] Neurocognition and Action–Biomechanics Research Group, Faculty of Psychology and Sports Science, Bielefeld University, 33615 Bielefeld, Germany; thomas.schack@uni-bielefeld.de
[3] Department of Health, Physical Education and Recreation, University of Cape Coast, Cape Coast PMB TF0494, Ghana
* Correspondence: s.simmons1@lse.ac.uk (S.S.S.); elvis.hagan@ucc.edu.gh (J.E.H.J.)

Abstract

Studies on the demography and health of racially diverse African populations are scarce, particularly due to lingering data challenges. Generative data modelling has emerged as a valuable solution to this burden. The study, therefore, examined the efficacy of Conditional Tabular GAN (CTGAN), CopulaGAN, and Tabula Variational Autoencoder (TVAE) for generating synthetic but realistic demographic and health data. This study employed the World Health Organisation stigy on global ageing and adult health survey (SAGE) Wave 1 South African data (n = 4227). Information missing from SAGE Wave 1, including demographic (e.g., race, age) and health (e.g., hypertension, blood pressure) indicators, were imputed using Generative Adversarial Imputation Nets (GAIN). CopulaGAN, CTGAN, and TVAE, sourced from the sdv 1.24.1 python library, generated 104,227 synthetic records based on the SAGE data constituents. The outcomes were accessed with similarity and machine learning (XGBoost) augmentation metrics (sourced from the sdmetrics 0.21.0 python library), including column shapes and overall and precision ratio scores. Generally, the GAIN imputations resulted in data with properties that were comparable to original and with no missing information. CTGAN's (89.20%) overall quality of performance was above that of TVAE (86.50%) and CopulaGAN (88.45%). These findings underscore the usefulness of generative data modelling in addressing data quality challenges in diverse populations to enhance actionable health research and policy implementation.

Keywords: deomography; synthetic data; generative data modelling; GAIN; CopulaGAN; CTGAN; TVAE; South Africa; Africa

Academic Editor: Eric Pardede

Received: 25 May 2025
Revised: 25 June 2025
Accepted: 14 July 2025
Published: 17 July 2025

Citation: Simmons, S.S.; Hagan, J.E., Jr.; Schack, T. Generative Data Modelling for Diverse Populations in Africa: Insights from South Africa. *Information* **2025**, *16*, 612. https://doi.org/10.3390/info16070612

1. Introduction

Africa has one of the most diverse populations in the world, characterised by genetic, cultural, and phenotypic diversity, as well as more than 2000 ethno-linguistic groups [1]. Furthermore, the individual countries on the continent have inherently diverse population characteristics. For instance, Nigeria and Uganda have about 371 [2] and 65 distinct ethnic divisions [3], respectively. Similarly, South Africa and Mauritius rank among the most racially diverse countries on the continent [4]. While ethnic differentials manifest in behavioural patterns, racial differentials encompass biological traits and socio-cultural behaviours. The interplay between genetics (biology) and socio-cultural factors influences health more than behavioural patterns only [5]. Thus, racial diversity, which entails ethnic diversity, across the continent may be more important in driving health outcomes, disease

patterns, and therapeutic responses [6,7]. Additionally, race and associated genetic and social determinants interact to influence disease burden. In this context, race may be a precursor that shapes exposure to other social indicators that modulate disease burden and health patterns [8].

Various types of data on demographic and health characteristics of racially diverse African populations exist, with tabular datasets comprising the majority [9,10]. These tabular datasets, particularly those from this century, exhibit distinct age-related health patterns. Young people are susceptible to infectious diseases and associated drivers, while adults face a higher risk of non-communicable diseases (NCDs), especially cardiometabolic diseases [11–13]; stroke, angina pectoris (angina), hypertension and diabetes are among the most prevalent forms of cardiometabolic diseases in adults [14–17]. In addition, the data reveals that the burden of these diseases is further polarised across various social determinants of health, with race/ethnicity, wealth, education, and residence as the most recurring factors [4,18]. The insights from such data underscore the significance of demographic and health data as indispensable for informed decision-making mechanisms in public health, policy, planning, and resource allocation within countries. However, the ethnic and/or racial diversity is often underrepresented in many health research reports, resulting in underreporting of population-specific demographic and health insights. Research by Naz et al. [19] and Kinyondo and Pelizzo [20] revealed that poor-quality data rank among the leading factors for the endemic underreporting of health insights in racially diverse countries in Africa. Indeed, the continent has limited funds for data collection. Accordingly, data collection efforts in most African countries rely on external funds and are typically carried out through large-scale surveys or national studies. Such studies have multiple constraints, including smaller sample sizes, missing data, privacy concerns, and restricted access due to ethics and regulations. Similarly, the archiving and management of the acquired data are suboptimal. Repositories are usually inaccessible and characterised by poor database preservation practices, collectively degrading data quality and limiting the utility of such datasets [19,21]. In addition, these limitations impede the effective use of demographic and health information for advanced analytics and the design of effective health promotion interventions for diverse populations in Africa. Another aspect is that the measurable reports on demography and health in racially diverse populations are based on obsolete, small samples with missing and incomplete data. For instance, Pillay-van Wyk et al. [4] report on one of the most racially diverse countries in Africa, South Africa, revealed variations in age-standardised deaths among different racial groups. Though the study was published in 2016, the data for analysis were four years older (1990–2012). Also, Mhlanga and colleagues' [22] study on racial differences in public healthcare revealed significant differences in healthcare by race. Their study, conducted by means of a survey in 2018, employed a sample size of 20,908 drawn from a population of 58.61 million. While that survey might have been representative of the population, a larger sample size may yield a more representative and reliable outcome [23]. As the challenge persists, it remains unclear whether data quality can be systematically improved in Africa, particularly for diverse populations. A practical and sustainable approach to improving the quality of demographic and health data would include minimising missing information, increasing data size, and minimal privacy concerns. Such an initiative will enhance the efficiency and impact of research work in African countries with diverse populations.

In recent years, deep learning techniques have been resurgent as prominent tools for overcoming challenges of data availability, accessibility, and quality in countries [24–27]. Indeed, using deep learning techniques to produce data that imitates real-world information without compromising privacy or having counterintuitive patterns will offer valuable opportunities for advancing research. Among the many deep learning techniques, generative

adversarial networks (GANs) and variational autoencoders (VAE) emerge as high-quality approaches for synthetic data generation [25,28]. Moreover, GANs and VAE are the foundational architectures for generative models tailored to tabular, text, and image data [29,30]. Accordingly, several GANs and VAE techniques have outperformed many methods in creating data similar to original tabular datasets. In racially diverse populations in Africa, data synthesis with GANs and VAEs could unlock new opportunities for cross-institutional research and predictive modelling without compromising ethical standards and responsible data practices. However, the application of GANs and VAEs to address the suboptimal quality of current demographic and health data from racially diverse populations in Africa remains largely underexplored. Consequently, there is uncertainty about whether GANs, VAEs, or both can be practical tools for generating data that can capture the statistical characteristics and correlations of a real-life survey dataset from racially diverse African populations. This gap warrants the utility of GANs and VAEs for synthetic data generation in racially diverse African populations. Evaluating the performance of GANs and VAEs in generating synthetic tabular data will demonstrate the potency of synthetic data in overcoming existing data access and quality barriers in Africa, enabling a more robust analysis of health and demographic trends. Additionally, such an assessment will provide information on the comparative evaluation of deep learning models' performance in health and demographic information generation based on a real-world survey from the African context.

Therefore, the study examined generative data modelling for diverse populations in Africa. Specifically, this study addresses the following primary research questions: (1) To what extent can generative data modelling techniques, including TVAE, CTGAN, and CopulaGAN, capture the statistical characteristics and correlations of diverse population data? (2) Which generative data modelling technique produces the highest-quality synthetic data in a diverse population?

2. Materials and Methods

2.1. The Context

South Africa, an ethnically and racially diverse country with known racial health disparities [4,22], is a suitable context for sourcing population and health data as baseline information for generative data modelling. The country's population comprise four main racial groups, namely, black (African [81.7%]), Indian (Asian [8.5%]), coloured (mixed race [2.6%]), and white (7.2%) [31]. These racial groups have varied ethnic identities and different socio-demographic statuses and lifestyles. The country, like most diverse populations in low-resource settings, is experiencing a rapid epidemiological shift marked by increased incidence of unhealthy weight gain and chronic disease, particularly in adults [16,32]. Furthermore, the rising incidence of mortality and morbidity associated with cardiometabolic conditions like ischaemic heart disease (IHD), stroke, and diabetes has accompanied the nutritional and epidemiological transition in the country [4]. South Africa, therefore, presents a valuable study setting for this research.

2.2. Data and Data Source

This study used data from the World Health Organisation (WHO) Global Ageing and Adult Health survey (SAGE) for South Africa. Due to access limitations for the most recent SAGE waves (2 and 3), this study was granted access to the South African SAGE (Wave 1). SAGE Wave 1 is the second wave of a longitudinal demographic and health study of persons aged 50 and above in six low- and middle-income countries (LMICs), including South Africa, conducted from 2007 to 2010. SAGE uses multistage cluster sampling techniques to provide information for 4227 South Africans [11,33]. The data capture several population

and health indicators, including socio-demographic characteristics, health state, biomarkers, anthropometric indices, risk factors, and chronic conditions relevant to the current study.

2.3. Variables and Variable Transformation

To further demonstrate the utility of deep learning for improving data quality in a diverse population, demographic and health variables were selected from the South SAGE Wave 1. Specifically, the chosen variables included age, sex, education, wealth, weight, height, blood pressure (BP), waist circumference (WC), and self-reported diagnoses of diabetes, angina pectoris (angina), stroke, and hypertension. Ever been to school, the education indicator, was recoded as either yes or no. Wealth was recoded as 'rich' or 'not rich'. These variables were selected because of their known association with the ongoing disease transition in South Africa and other diverse populations in Africa [4,11,12,15,16]. Weight values outside the range of 50 kg to 300 kg were set to not available (NA). Height values out of range 80 cm and 300 cm were set as NA. The WC values out of the range 50 cm to 300 cm, were set to NA. BP values out of the range 65 mmHg and 300 mmHg were set to NA. Self-reported disease status and educational (ever been to school) statuses other than 'yes' or 'no' were set to NA. The recoding of these variables is because human biomarkers and anthropometric measurements fall within particular ranges. Also, self-reported health status is often measured on a binary scale, i.e., whether the subject has (yes) or does not have (no) the condition. Thus, values below or above the threshold for weight, height, WC, and disease status are rare and may indicate errors of measurement or data entry. Setting these to NA prevented unrealistic values from skewing analyses. Moreover, including these extreme values in the analysis would result in weak statistics [11]. On top, these transformations enabled further computation of the missing information using the deep learning techniques.

2.4. Handling Missing Data

The data were characterised by missing information. Accordingly, the study adopted the generative adversarial imputation nets (GAIN) method to impute all missing values. GAIN is a type of GANs typically designed to perform robust imputation on data with mixed features (categorical and numerical indicators). Moreover, GAIN can produce imputations that closely reflect the underlying distribution of the actual data [34]. The GAIN imputer builds upon the GAIN architecture to implement a robust deep learning-based approach for imputing missing values in tabular datasets from all settings, including the low-resource setting under study. The GAIN imputer was modified to emphasise statistical fidelity and safe convergence while addressing the challenge of missing information in the SAGE data. As indicated in Figure 1, this study followed several steps to create and apply the GAIN imputer to the missing information in the SAGE data.

First, the GAIN imputer was configured with hyperparameters to control specific portions of the model's behaviour, performance, and generalisability, thereby ensuring a reliable imputation. A fraction of the missing information was masked to a discriminator (D), measured at a hint rate of 0.9. The adversarial and reconstruction loss were balanced via a regularisation coefficient, alpha = 10, with a controlled training batch size = 64, and duration = 300 epochs [35]. These configurations did not contain an early stopping mechanism and were set to automatically select a graphical processing unit (GPU or compute unified device architecture [CUDA]) or central processing unit (CPU), depending on availability. Because the number of observations in the SAGE data was less than 10,000, this study employed a batch size ranging from 32 to 128. The hint rate was set at 0.9 to enable the discriminator (D) to more effectively locate missing information (almost complete masking). The epochs were reduced to 300 because some variables (e.g., hypertension and

stroke) had missing information ranging from 5% to 16%, and the overall dataset numbered approximately 1000 but was less than 10,000 [36–38].

Second, the GAIN imputer pre-processed the mixed data types extracted from the SAGE data. Here, individual data types were assigned different preprocessing methods, including min-max scaling and one-hot encoding, using the MinMaxScaler and OneHotEncoder from the scikit-learn 1.7.0 Python library. Numerical data columns, such as weight, height, and age, were normalised (0, 1) using the MinMaxScaler to mitigate the effects of values with high ranges on the expected outcomes. The categorical variables were encoded using a OneHotEncoder and converted to factors or numeric format (e.g., sex [male or female] became sex [0 or 1]). For all data types, missing values were tracked through a binary mask to guide the generator (G) during training. The generator (G) then received both data types (numerical and categorical) and the missingness mask. The G predicted missing values by outputting imputed values of coded and normalised range (0 or 1) using a sigmoid activation. Following the actions of G, a discriminator (D) was used to distinguish between observed and imputed values based on the data and a "hint" matrix, partially revealing the actual mask [34,39] (see Figure 1). The hint matrix revealed a portion of the missingness mask to the D, thereby preventing trivial identification of observed values and promoting more meaningful learning. The D also employed a sigmoid activation function because of its ability to map values between 0 and 1, thereby ensuring bounded gradients for all outputs [36,37]. An Adam optimiser was used to minimise the respective losses [34,40].

The third phase was postprocessing. Following the generation of imputed values, the regularised (normalised and label-encoded) features were decoded (for categorical indicators) and denormalised (for numerical indicators) to their original formats. These steps helped the reconstructed data to maintain its initial structure and statistical characteristics [34,36]. The loss values for both G and D were logged during the training phase to allow for output monitoring (see Figure 1). The GAIN imputation architecture and application code can be retrieved from "*Imputation with GAIN (code)*" on GitHub at https://github.com/SallySims/GAIN-Imputation-I.git (accessed on 24 June 2025).

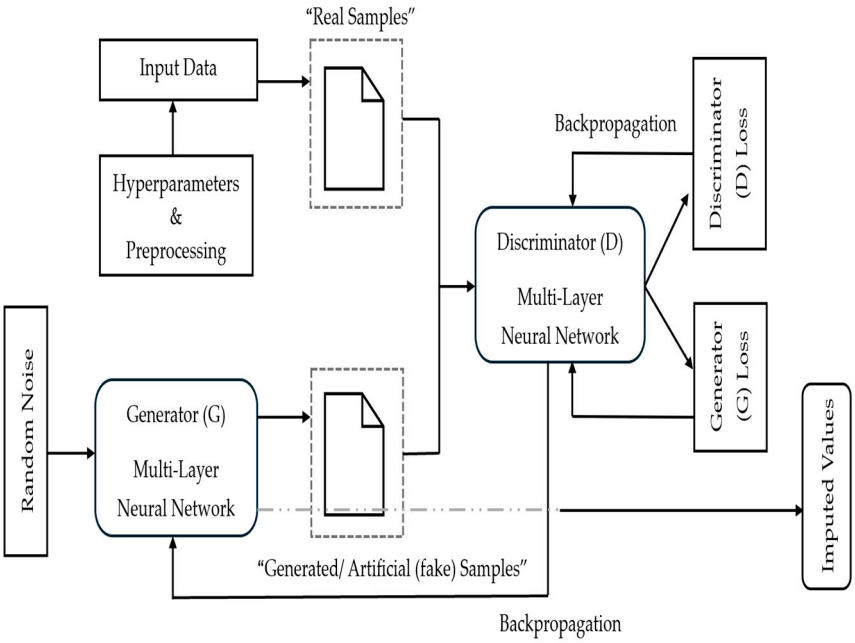

Figure 1. Generative Adversarial Imputation Network architecture diagram. Source: Authors' creation based on information from Yoon et al. [34], Patterson and Gibson [37].

To assess the similarity between the tabular data and the GAIN imputed data, statistical properties were evaluated using the Kolmogorov–Smirnov (KS) and Jensen–Shannon (JS) tests. These tests were selected due to their established strength in validating data distribution preservation metrics and are suitable for comparing convergence and divergence between original and imputed data [41]. Accordingly, the KS and JS tests assessed the outcomes for numerical and categorical variables, respectively.

2.5. Statistical Analysis

2.5.1. Descriptive Analysis

The original and imputed data were summarised, presented as means, standard deviation (SD), and percentages. Through these summaries, the distribution, the similarities, and differences between the variables in the original and imputed data are presented [12,42].

2.5.2. Inferential Analysis

This study employed CopulaGAN, CTGAN, and TVAE to perform generative data modelling due to the need for diverse and increased response volumes in South Africa's diverse population. These GANs and VAE were selected because of their superior performance in generating synthetic tabular information for datasets with mixed features, in contrast with machine learning, and traditional approaches [27,30]. The CTGAN model efficiently employs mode-specific normalisation processes, builds on a conditional generator and training-by-sampling, and uses fully connected networks and other contemporary data manipulation techniques to create high-quality data synthesis outputs [43–46]. Specifically, the CTGAN employs mode-specific normalisation for the independent preprocessing of each variable in all data types. In this method, a variational Gaussian mixture model (VGM) estimates each continuous column's modes, because such columns often have non-Gaussian multimodal distributions. Each value is then represented with two components: a one-hot vector indicating the mode, and a scalar showing the value of the mode. Furthermore, the CTGAN introduces a conditional generator to condition sample generation on a selected category from a discrete column. This approach addresses the challenges of standard GAN generators, particularly their tendency to overlook category imbalance, which leads to poor representation of infrequent categories. The condition is represented by a vector that combines one-hot encodings and mask vectors to specify the targeted category. Generated outputs are penalised via cross-entropy loss if the outputs fail to match the condition. This ensures the model learns actual conditional distributions. On top, the CTGAN employs a training-by-sampling approach to enhance balanced representation during training. A discrete column is randomly selected, and a category is sampled based on the logarithm of its frequency. This guides the generator (G) to create realistic samples for the discriminator (D) to evenly explore all categories as either real or fake, while maintaining the original data distribution. Both the generator (G) and the discriminator (D) use two connected layers. While the G applies batch normalisation with Rectified Linear Unit (ReLU), tanh, and Gumbel-Softmax activation functions, the D uses leaky ReLU and dropout. CTGAN employs the Packed Generative Adversarial Networks (PacGAN) framework to mitigate mode collapse and utilises the Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) loss, combined with the Adam optimiser. These and specific hyperparameter scores (epochs, batch size, embedding dimension, generator (G) and discriminator/critic (D), and G and D learning rate, see https://github.com/sdv-dev/SDV/tree/main/sdv/single_table (accessed on 24 June 2025) for details) are introduced as part of the training to improve the model's performance. A comprehensive explanation of the CTGAN framework is reported by Xu et al. [46] and Precise et al. [45].

The CopulaGAN is a variant of CTGAN that leverages a copula-based modelling technique for data synthesis [47]. The CopulaGAN model uses the Gaussian copulas to analyse non-normal dependent observations by modelling dependency through a multivariate normal error structure to separate the dependence structure from the marginal distributions [48]. Furthermore, the copulas apply cumulative distribution function (CDF)-based transformations to data to simplify the learning of data trends. Inference within this model relies on maximum likelihood, with closed-form likelihoods for continuous data and numerical approximations (e.g., importance sampling) for discrete/categorical data [30,47–49]. Like CTGAN, hyperparameters (epochs, batch size, embedding dimension, generator (G) and discriminator/critic (D), and the learning rates of G and D) are introduced as part of the data synthesis process. Masarotto and Varin [48] and Pathare and colleagues [47] provide detailed information on Gaussian copulas and CopulaGAN.

TVAE employs two neural networks and a conditional probabilistic distribution to generate artificial data. Within this architecture, a similar CTGAN preprocessing step is implemented, although the loss function is modified to better suit data with a tabular structure. The TVAE use an encoder–decoder and Evidence Lower Bound (ELBO) loss to train a model [30]. One of the two neural networks, the decoder, models the condition distribution, whereas the other, the encoder, models the approximate posterior. Specifically, the decoder outputs a joint distribution over continuous (2Nc) and discrete (Nd) indicators. Here, the Nd is modelled with Gaussian distributions while the discrete values and mode indicators are modelled with categorical distributions via Softmax. On the one hand, the encoder maps an input row to latent indicators, and on the other hand, it employs standard VAE encoding with outputs for the mean and log variance, while assuming the latent indicator is normally distributed. Also, the model is trained using an Adam optimiser and other hyperparameters [30,46,50]. More detailed information on TVAE architecture is presented by Xu et al. [46].

This study utilised synthetic data vaults (SDVs) project algorithms presented as sdv 1.24.1 python library to apply TVAE, CopulaGAN, and CTGAN to augment the original South African SAGE demographic and health data. The data synthesis process using SDV was built on the standard approach for generative data modelling. Additionally, the SDVs incorporated parsimonious yet statistically significant hyperparameters into the data augmentation procedure [51,52]. For instance, the CTGAN and CopulaGAN hyperparameter configurations included epochs = 300, batch size = 500, embedding dimension = 128, generator (G) and discriminator/critic (D) = 256 and 256, and G and D learning rate = 2×10^{-4}. Information on the details of SDV data synthesis for the three generative models can be found at https://github.com/sdv-dev/SDV/tree/main/sdv/single_table. These generative modelling techniques can generate data at about 1 to 100 times the initial scale. Therefore, this study set the synthetic data output to 104,227, i.e., ~25 $\times$ 4227 (the sample size). Also, the synthetic data metrics (SD metrics) were used to evaluate how closely the synthetic data generated with the TVAE, CopulaGAN, and CTGAN compared with the original data in terms of complexities, dependencies, and characteristics, i.e., the quality of reporting. The SD metrics, from the scikit learn 1.7.0 python library, was used to evaluate the distribution of columns (column shapes) in the synthetic and real data, using similarity metrics such as the Jensen–Shannon (JS) divergence (for categorical indicators) and the Kolmogorov–Smirnov (KS) test (for numerical indicators). Specifically, the KS test compared the empirical cumulative distribution functions of the CTGAN, CopulaGAN, and TVAE generated data with the original dataset. For this statistic, KS test, the maxi-

mum value (supremum [sup_x]) of the absolute difference between the original $F_r(n)$ and generated $F_g(n)$ data is of interest. This is presented in Equation (1) as follows:

$$D = sup_x \left| F_r(n) - F_g(n) \right| \tag{1}$$

The derived KS statistic is converted to a KS score by normalisation as stated in Equation (2).

$$KS_{score} = (1 - D) \tag{2}$$

The JS statistic examines the divergence (zero divergence or symmetry) $DJS(R \parallel S)$ between the real (R) and synthetic (S) data distributions based on the Kullback–Leibler divergence (KLD), as presented in Equation (3)

$$DJS(R \parallel S) = \frac{1}{2} KLD(R \parallel M) + \frac{1}{2} KLD(S \parallel M) \tag{3}$$

where the following apply:

M = $\frac{1}{2}(R + S)$ is the average (or mixture) of the two distributions (real and generated).

$DJS(R \parallel S)$ = 0 if and only if R is exactly the same as S. However, small DJS indicate symmetry.

$KLD(R \parallel M)$ is the KLD from R to M, defined in Equation (4) as follows:

$$KLD(R \parallel M) = \sum_i R(i) log(\frac{R(i)}{M(i)}) \tag{4}$$

Like KS, the JS output $DJS(R \parallel S)$ is converted to a normalised DJS_{score} (range 0 to 1), as presented in Equation (5):

$$DJS_{score} = 1 - DJS(R \parallel S) \tag{5}$$

The column shape metric averages KS and JS scores for columns within specific variable types (numerical [$i \epsilon C_{num}$] and categorical [$i \epsilon C_{cat}$]), defined as Equation (6):

$$Column\ Shape\ Metric = \frac{1}{n} (\sum_{i \epsilon C_{cat}} DJS_{score,i}) + (\sum_{i \epsilon num} KS_{score,i}) \tag{6}$$

Also, the pairwise relationships between the features (column pair trends) were evaluated for consistency. This ensured that the synthetic data captured both the individual column distributions (univariate properties) and relationships between the columns (multivariate structure) in the original data. Based on the SD metric specifications, the two main metrics of correlation and contingency similarity were employed for column pair trend analysis [51,52]. The column pair trend metrics compute the correlation coefficient (either Pearson or Spearman) for a pair of columns, A and B, in the original (R) and synthetic (S) data. This yields two separate correlation values. The test normalises and returns a similarity score based on Equation (7):

$$score = 100 \times \left(1 - \frac{|S_{A,B} - R_{A,B}|}{2} \right) \tag{7}$$

For a pair of columns, A and B, the contingency similarity test computes a normalised contingency table for the real (R) and synthetic (S) data. The contingency table describes the proportion of rows with each combination of categories in columns A and B. Furthermore, the test uses the total variation distance ($\frac{1}{2} \sum_{\alpha \epsilon A} \sum_{\beta \epsilon B} \left| S_{\alpha,\beta} - R_{\alpha,\beta} \right|$) approach to compute

the difference between the contingency tables. The derived distance is subtracted from 1, ensuring a high score means high similarity. The process is summarised in Equation (8):

$$score = 100 \times \left(1 - \frac{1}{2}\sum_{\alpha \epsilon A}\sum_{\beta \epsilon B}\left|S_{\alpha,\beta} - R_{\alpha,\beta}\right|\right) \quad (8)$$

The average of all metric outcomes became the overall score. All the outputs were presented in percentages to provide a universal and intuitive scale for the quality measurements [53,54]. Machine learning (ML) augmentation metrics were used in the multivariate analysis to test the extent to which the dual datasets' multivariate structures yielded better results than the original data when training a model. Regarding the SD metric, this study used the 'binary classifier precision efficacy' method to analyse how precisely (precision score) the classifier (XGBoost, sourced from the sdmetrics 0.21.0 python library) trained on synthetic data generated from the real data. XGBoost was used because of its ability to handle large datasets [55]. Here, regarding sensitivity, the current study employed hypertension, the condition with the most reported and derived actual cases, as the prediction outcomes of interest at a fixed recall level of 0.8. Heart conditions like hypertension and angina pectoris or ischaemic heart disease (IHD) often receive general risk scoring with a target recall of about 0.7–0.85. Hence, setting the recall at 0.8 aligns with medical diagnosis standards that prevent missing too many actual cases of diseases or conditions of interest [56,57]. All feature engineering and analyses were performed with Python 3.1.25.

3. Results

Table 1 illustrates a comparative summary of the demographic and health characteristics in the original and imputed datasets for 4227 South African adults. Generally, there were differences in the primary variable outcomes, but similarities in the outputs of the variable elements. More numerical than categorical indices recorded changes in outcomes following the imputation with GAIN. In the original dataset, the average age of a South African was 62.71 ± 9.65 years, whereas the modified data revealed a decreased average age of 62.63 ± 9.83 years. Average height increased from 157.95 ± 12.52 cm to 158.55 ± 13.19 cm. Systolic blood pressure (BPs) decreased slightly from 145.34 ± 25.36 mmHg to 145.00 ± 25.20 mmHg. Racial status was characterised initially by a missing value (643 (15.21) and other statuses (African/Black: 2238 (52.95%), Coloured: 716 (16.94%), Indian/Asian: 335 (7.93%), White: 287 (6.79%), Other: 8 (0.19%)). Hypertensive status was reported among 1144 (27.06%) respondents, although 204 (4.83%) provided no response. Following the imputation with GAIN, racial status distribution was as follows: African/Black: 2344 (55.45%), Coloured: 822 (19.45%), Indian/Asian: 576 (13.63%), White: 477 (11.28%) and Other: 8 (0.19%). The consistencies in statistical properties ($D_{n,m}$ (KS)) in both datasets ranged from 0.01 for BPs to 0.02 for the other variables (age, height, weight, WC), at $p > 0.05$ (no statistically significant difference). Hypertension status report was devoid of NA/no response and increased to 1232 (29.15%)

Table 2 illustrates the outcomes of generative data modelling using TVAE, CTGAN, and CopulaGAN, sourced from the sdv 1.24.1 python library. A total of 104,227 records were generated by the three models, even though the outcomes varied according to the variables under study. These generative data models provided information on demographics and health indicators, including race, age, sex, BPs, self-reported angina, height, and wealth. Comparing the numerical variables, the average age of participants was oldest for CTGAN (63.30 ± 9.87 years), older for CopulaGAN (63.04 ± 12.85 years), and old for TVAE (59.26 ± 11.70 years). The average BP generated from CTGAN (142.17 ± 25.58 mmHg) was within the average estimates from TVAE (138.30 ± 23.25 mmHg) and CopulaGAN (145.81 ± 28.17 mmHg). The average weight and WC deduced from CopulaGAN [weight

(76.86 ± 15.41 kg) and WC (94.23 ± 21.64 cm)] were higher, in contrast with the averages from TVAE [weight (72.95 ± 15.65 kg) and WC (91.14 ± 17.65 cm)], which were lower. Regarding categorical indicators, there were more African/Black racial identities than other groups in all models. CTGAN produced a relatively balanced race distribution (i.e., 51.82% African/Black, 20.90% coloured, 18.10% Indian/Asian, 8.19% White, 0.99% Other) compared to TVAE (72.69% African/Black, 10.19% coloured, 5.50% Indian/Asian, 11.62%White, NA for other races). Females were more common than males across all generative data models, although CTGAN had more males [49,398 (47.39%)] than TVAE (37,121 (35.62%)) or CopulaGAN 42.65%, n = 44,452). The generated prevalence of hypertension varied between the models, with CTGAN [38,392 (36.83%)] recording the highest and TVAE [24,281 (23.30%)] revealing the lowest prevalence. A similar pattern was observed for stroke and diabetes and angina pectoris.

Table 1. Summary of the Original and Imputed Data.

	Original	**Imputed**	**Tests**
	Number (%) 4227 (100) $\bar{x} \pm$ SD	Number (%) 4227 (100) $\bar{x} \pm$ SD	stats, *p*-value $D_{n,m}$ (KS)
Age	62.71 ± 9.65	62.63± 9.83	0.02, $p > 0.05$
Height	157.95 ± 12.52	158.55 ± 13.19	0.02, $p > 0.05$
Systolic blood pressure (BPs)	145.34 ± 25.36	145.00 ± 25.20	0.01, $p > 0.05$
Weight	76.44 ± 18.30	76.08 ± 19.72	0.02, $p > 0.05$
Waist circumference (WC)	94.40 ± 17.44	95.66 ± 19.00	0.02, $p > 0.05$
	Number (%)	Number (%)	DJS (P‖Q) (JS)
Sex			0.01
Male	1797 (42.51)	1797 (42.51)	
Female	2428 (57.45)	2430 (57.49)	
Missing	2 (0.047)	NA	
Race			0.05
African/Black	2238 (52.95)	2344 (55.45)	
Coloured	716 (16.94)	822 (19.45)	
Indian/Asian	335 (7.93)	576 (13.63)	
White	287 (6.79)	477 (11.28)	
Other	8 (0.19)	8 (0.19)	
Missing	643 (15.21)	NA	
Ever been to school			0.01
Yes	2661 (62.93)	3043 (71.99)	
No	873 (20.65)	1184 (28.01)	
Missing	693 (16.41)	NA	
Wealth			0.01
Rich	1767 (41.80)	2123 (50.22)	
Not rich	1638 (38.75)	2104 (49.78)	
Missing	822 (19.45)	NA	
Angina pectoris			0.02
Yes	229 (5.42)	351 (8.30)	
No	3798 (89.88)	3876 (91.70)	
Missing	200 (9.53)	NA	
Hypertension			0.01
Yes	1144 (27.06)	1232 (29.15)	
No	2879 (68.11)	2995 (70.85)	
Missing	204 (4.83)	NA	

Table 1. *Cont.*

	Original	Imputed	Tests
Stroke			0.05
Yes	144 (3.41)	340 (8.04)	
No	3883 (91.86)	3887 (91.96)	
Missing	200 (4.73)	NA	
Diabetes			0.02
Yes	370 (8.75)	453 (10.72)	
No	3657 (86.52)	3774 (89.28)	
Missing	200 (4.73)	NA	

Source: Computation based on data from WHO SAGE South Africa [33]. Note: DJS (P‖Q) (JS), Jensen–Shannon divergence; $D_{n,m}$ (KS), Kolmogorov–Smirnov; NA, not available.

Table 2. Summary of the synthetic data generated with CTGAN and CopulaGAN.

	CopulaGAN	CTGAN	TVAE
	Number (%)	Number (%)	Number (%)
	104,227 (100)	104,227 (100)	104,227(100)
Indicators	$\bar{x} \pm$ SD	$\bar{x} \pm$ SD	$\bar{x} \pm$ SD
Age	63.04 ± 12.85	63.30 ± 9.87	59.26 ± 11.70
Height	158.43 ± 16.76	159.86 ± 19.21	159.07 ± 9.91
Systolic blood pressure (BPs)	145.81 ± 28.17	142.17 ± 25.58	138.30 ± 23.25
Weight	76.86 ± 15.41	76.08 ± 19.68	72.95 ± 15.65
Waist circumference (WC)	94.23 ± 21.64	92.82 ± 20.08	91.14 ± 17.65
	Number (%)	Number (%)	Number (%)
Sex			
Male	44,452 (42.65)	49,398 (47.39)	37,121 (35.62)
Female	59,775 (57.35)	54,829 (52.61)	67,106 (64.38)
Race			
African/Black	47,764 (45.83)	54,012 (51.82)	75,760 (72.69)
Coloured	25,866 (24.82)	21,782 (20.90)	10,619 (10.19)
Indian/Asian	13,822 (13.26)	18,868 (18.10)	5732 (5.50)
White	15,730 (15.09)	8535 (8.19)	12,116 (11.62)
Other	1045 (1.00)	1030 (0.99)	NA
Ever been to school			
Yes	77,156 (74.03)	84,540 (81.11)	85,914 (82.43)
No	27,071 (25.97)	19,687 (18.89)	18,313 (17.57)
Wealth			
Rich	49,919 (47.89)	47,907 (45.96)	43,095 (41.35)
Not rich	54,308 (52.11)	56,320 (54.04)	61,132 (58.65)
Angina pectoris			
Yes	20,770 (19.93)	16,309 (15.65)	419 (0.40)
No	83,457 (80.07)	87,918 (84.35)	103,808 (99.60)
Hypertension			
Yes	29,595 (28.39)	38,392 (36.83)	24,281 (23.30)
No	74,632 (71.61)	65,835 (63.17)	79,946 (76.70)
Stroke			
Yes	14,446 (13.86)	20,543 (19.71)	1684 (1.62)
No	89,781 (86.14)	83,684 (80.29)	102,543 (98.38)
Diabetes			
Yes	13,074 (12.54)	23,158 (22.22)	2116 (2.03)
No	91,153 (87.46)	81,069 (77.78)	102,111 (97.97)

Source: Computation based on data from WHO SAGE South Africa [33].

Table 3 shows the assessment of the quality of generative data modelling with CopulaGAN, CTGAN, and TVAE, focusing on how these three models preserved the original data's complex interrelationships and statistical properties. These quality reports were

presented as column shapes, pair trends, and overall scores. The CTGAN outputs were of the highest quality (89.20%) compared with TVAE (86.50%), which were the lowest. Also, CTGAN (90.16%; 88.25%) outperformed TVAE (89.86%; 83.13%) and CopulaGAN (89.63%; 87.27%) in terms of column shapes and column pair trends outcomes.

Table 3. Generative data model quality metrics (SD metrics) for CopulaGAN, CTGAN, and TVAE.

Metric	CopulaGAN	CTGAN	TVAE
Column Shapes	89.63%	90.16%	89.86%
Column Pair Trends	87.27%	88.25%	83.13%
Overall Score	88.45%	89.20%	86.50%

Source: Computation based on data from WHO SAGE South Africa [33].

Table 4 shows the ML augmentation metrics outcomes for the CTGAN, CopulaGAN, and TVAE generated data and the original information for South Africa. The metric compares the performance of these three models' training and validation dataset, focusing on recall, precision, and confusion matrix (TP, FP, TN, FN), along with a precision ratio score. According to Table 4, all the synthetic data significantly improved recall (validation recall) [COpulaGAN: 0.47, CTGAN:0.47, TVAE:0.44] compared with the original data (0.23). The original data were more precise (0.59) than the synthetic datasets (CopulaGAN: 0.45, CTGAN: 0.46, TVAE: 0.43) at correctly predicting positive hypertensive cases (validation precision). The ratio of correctly predicted positive cases of hypertension (precision ratio score) in the synthetic versus original data revealed that CTGAN (0.77) and CopulaGAN (0.76) achieved higher scores than the TVAE model (0.73).

Table 4. Generative data models' ML augmentation metrics (SD metrics) for CopulaGAN, CTGAN, and TVAE.

	CopulaGAN		CTGAN		TVAE	
	Data Types					
	Original	Synthetic	Original	Synthetic	Original	Synthetic
Training Recall	0.80	0.80	0.80	0.80	0.80	0.80
Validation Recall	0.23	0.47	0.23	0.47	0.23	0.44
Validation Precision	0.59	0.45	0.59	0.46	0.59	0.43
True Positives (TPs)	82	168	82	169	82	168
False Positives (FPs)	56	205	56	201	56	220
True Negatives (TNs)	852	703	852	707	852	688
False Negatives (FNs)	279	193	279	193	279	210
Score (Precision Ratio)	1.0 (baseline)	0.76	1.0 (baseline)	0.77	1.0 (baseline)	0.73

Source: Computation based on data from WHO SAGE South Africa [33].

4. Discussion

This study presents a pioneering report on applying generative data modelling techniques to improve data quality in diverse populations across Africa. Specifically, this study employed CopulaGAN, CTGAN, and TVAE for generating synthetic data for South Africa, a racially and ethnically diverse African population. These techniques mimicked the complex dependencies and statistical relationships in the WHO SAGE South Africa Wave 1 (4227 observations) data to enhance synthetic data generation (104,227 observations). The quality of outputs from these generative data modelling techniques revealed differential levels of similarity between original and artificial data. These findings underscore the importance of generative data modelling as an efficient mechanism for overcoming data quality challenges and capturing the dynamics in diverse populations in Africa.

The revelation that GAIN, CTGAN, CopulaGAN and TVAE mimicked the statistical and complex dependency characteristics of the WHO SAGE South African data to impute missing information and generate data at a scale 24.66 times the original data highlights the usefulness of generative data modelling for mitigating issues of poor data quality for diverse populations. The data generation capabilities of these models corroborate reports [27,29,30,34,58] of high-accuracy imputations for missing information in mixed data types. Additionally, these studies report the effectiveness of these models in creating privacy-preserving synthetic data across different population groups and clinical contexts, supporting informed health decisions. These findings underscore the transformative potential of these models as scalable solutions for enhancing data quality in diverse populations and underserved areas, especially in Africa, where data challenges are widespread.

The observed variations in the SD metrics outcomes for CTGAN, CopulaGAN, and TVAE could be attributable to many factors. It is possible that the relatively strong performance of CTGAN compared to the other models may be due to the structure of the variables selected from the South African data. CTGAN performed well in learning relationships among categorical indicators, rendering it well-suited to modelling synthetic data with many such indicators, such as the variables employed in the current study. In addition, CTGAN uses mode-specific normalisation and conditional vector sampling for data synthesis. These techniques allow CTGAN to model discrete variables more efficiently than CopulaGAN and TVAE [27,29,43]. In contrast, the poor performance of TVAE relative to the GAN models (CopulaGAN and CTGAN) may have been due to its limited ability to model non-linear relationships in tabular data. Similarly, TVAE employs one-hot encoding to preprocess categorical data, transforming the data into binary vectors; however, it may not capture complex relationships or handle data with a significant class imbalance. The reasonably commensurate outcomes between CTGAN and CopulaGAN might be due to the similarity in their mode of synthetic tabular generation architecture [30,50].

4.1. Implications for Low-Resource Settings

The current findings highlight the suitability of generative data modelling techniques for application to diverse populations. By generating artificial data that exceeds the scale of the original data by a factor of two, these models can provide cost-effective and ethically sound data generation for advanced predictive analytics. In the context of Africa, where diverse populations are experiencing rapid epidemics of NCDs [4,15], CTGAN and CopulaGAN can become valuable mechanisms for generating data on the racial distribution of NCDs and associated drivers. Access to such data for advanced predictive modelling will therefore be crucial for informing evidence-based health policy built on context-specific, yet less expensive, data. Moreover, because TVAE and CTGAN have peculiar methodological advantages in the synthesis of data, these two models can be combined for more reliable data generation outcomes.

4.2. Limitations and Future Directions

Despite the strength of the study, some limitations can be noted. The study relied solely on WHO SAGE South Africa Wave 1 for analysis, and the data may not fully represent population diversity in Africa. Data from other racially and ethnically diverse African areas should be incorporated into future studies. Also, the study used only two forms of GANs, CTGAN and CopulaGAN, despite the existence of multiple forms of GANs designed for tabular data. However, CTGAN and CopulaGAN used in the study are models that have been identified in the literature as effective for the production of realistic and equitable synthetic data outcomes. The analysis also revealed that the models performed well with specific aspects of the data. Hence, future research could integrate these models to facilitate

better learning from the data and the subsequent execution of improved synthetic data generation outcomes.

5. Conclusions

Generative data modelling techniques can address the lingering data limitations and challenges hindering the complete execution of actionable research on diverse African populations. Our findings suggest that each model, CTGAN, CopulaGAN, and TVAE, performed well in specific aspects of synthetic data generation. The CTGAN scored the most balanced representation across variables, especially in racial classes. However, the overall performance of the CopulaGAN was below that of the CTGAN; the former better preserved inter-variable relationships. Also, the overall performance of TVAE was lower than that of CTGAN. TVAE exhibited stronger performance outputs for variables such as BPs and age, which had non-linear relationships. Therefore, generative data models offer a valuable pathway to more representative research and policy-making. However, further validation studies are needed to assess the external utility and long-term implications of using such data for African health research.

Author Contributions: Conceptualisation, S.S.S.; methodology, S.S.S.; formal analysis, S.S.S.; data curation, S.S.S. and J.E.H.J.; writing—original draft preparation, T.S., S.S.S., and J.E.H.J.; writing—review and editing, S.S.S. and J.E.H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used for the study are available at: https://apps.who.int/healthinfo/systems/surveydata/index.php/catalog/5 (accessed on 2 June 2025).

Acknowledgments: The authors thank Tiziana Leone and Grace Lordan for their insightful comments.

Conflicts of Interest: The authors declare that there are no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

2Nc	Continuous variables
BPs	Systolic blood pressure
CDF	Cumulative distribution function
CopulaGAN	Copula generative adversarial network
CTGAN	Conditional tabular generative adversarial network
D	Discriminator
DJS(P||Q)	Jensen–Shannon divergence
JS	Jensen–Shannon divergence
ELBO	Evidence lower bound
G	Generator
GAIN	Generative adversarial imputation nets
GAN	Generative adversarial network
FN	False negative
FP	False positive
KLD	Kullback–Leibler divergence
LMICs	Low- and middle-income countries
NCDs	Non-communicable diseases

Nd	Discrete variables
PacGAN	Packed generative adversarial network
ReLU	Rectified linear unit
SAGE	Study on global ageing and adult health
SD	Standard deviation
SD Metrics	Synthetic data metrics
SDV	Synthetic data vault
TN	True negative
TP	True positive
TVAE	Tabular variational autoencoder
VAE	Variational autoencoder
VGM	Variational Gaussian mixture model
WC	Waist circumference
WGAN-GP	Wasserstein generative adversarial network with gradient penalty
WHO	World health organisation

References

1. Campbell, M.C.; Tishkoff, S.A. African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annu. Rev. Genom. Hum. Genet.* **2008**, *9*, 403–433. [CrossRef] [PubMed]
2. Imoni, H.O. Exploring Ethnic Diversity on Managerial Choices in Nigeria. Ph.D. Thesis, Walden University, Minneapolis, MN, USA, 2018.
3. CIA Uganda—2022. World Factbook Archive. Available online: https://www.cia.gov/the-world-factbook/about/archives/2022/countries/uganda/ (accessed on 18 May 2025).
4. Pillay-van Wyk, V.; Msemburi, W.; Laubscher, R.; Dorrington, R.E.; Groenewald, P.; Glass, T.; Nojilana, B.; Joubert, J.D.; Matzopoulos, R.; Prinsloo, M.; et al. Mortality Trends and Differentials in South Africa from 1997 to 2012: Second National Burden of Disease Study. *Lancet Glob. Health* **2016**, *4*, e642–e653. [CrossRef] [PubMed]
5. Institute of Medicine (US) Committee on Assessing Interactions Among Social, Behavioral, and Genetic Factors in Health. The Impact of Social and Cultural Environment on Health. In *Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate*; Hernandez, L.M., Blazer, D.G., Eds.; National Academies Press: Washington, DC, USA, 2006.
6. Gomez, F.; Hirbo, J.; Tishkoff, S.A. Genetic Variation and Adaptation in Africa: Implications for Human Evolution and Disease. *Cold Spring Harb. Perspect. Biol.* **2014**, *6*, a008524. [CrossRef] [PubMed]
7. Liang, J.; Matheson, B.E.; Douglas, J.M. Mental Health Diagnostic Considerations in Racial/Ethnic Minority Youth. *J. Child Fam. Stud.* **2016**, *25*, 1926–1940. [CrossRef] [PubMed]
8. National Research Council (US) Panel on Race, Ethnicity, and Health in Later Life. *Understanding Racial and Ethnic Differences in Health in Late Life: A Research Agenda*; Bulatao, R.A., Anderson, N.B., Eds.; The National Academies Collection: Reports funded by National Institutes of Health; National Academies Press: Washington, DC, USA, 2004; ISBN 978-0-309-09247-0.
9. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *Inf. Fusion* **2022**, *81*, 84–90. [CrossRef]
10. Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S.B.; Schirrmeister, R.T.; Hutter, F. Accurate Predictions on Small Data with a Tabular Foundation Model. *Nature* **2025**, *637*, 319–326. [CrossRef] [PubMed]
11. Simmons, S.S. Strikes and Gutters: Biomarkers and Anthropometric Measures for Predicting Diagnosed Diabetes Mellitus in Adults in Low- and Middle-Income Countries. *Heliyon* **2023**, *9*, e19494. [CrossRef] [PubMed]
12. Simmons, S.S. Spatial Analysis of Metabolic Equivalents of Task Among Females in Urban and Rural Ghana. *Obesities* **2025**, *5*, 33. [CrossRef]
13. Simmons, S.S.; Hagan, J.E., Jr.; Schack, T. The Influence of Anthropometric Indices and Intermediary Determinants of Hypertension in Bangladesh. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5646. [CrossRef] [PubMed]
14. Allen, L.; Williams, J.; Townsend, N.; Mikkelsen, B.; Roberts, N.; Foster, C.; Wickramasinghe, K. Socioeconomic Status and Non-Communicable Disease Behavioural Risk Factors in Low-Income and Lower-Middle-Income Countries: A Systematic Review. *Lancet Glob. Health* **2017**, *5*, e277–e289. [CrossRef] [PubMed]
15. Bigna, J.J.; Noubiap, J.J. The Rising Burden of Non-Communicable Diseases in Sub-Saharan Africa. *Lancet Glob. Health* **2019**, *7*, e1295–e1296. [CrossRef] [PubMed]
16. Gouda, H.N.; Charlson, F.; Sorsdahl, K.; Ahmadzada, S.; Ferrari, A.J.; Erskine, H.; Leung, J.; Santamauro, D.; Lund, C.; Aminde, L.N.; et al. Burden of Non-Communicable Diseases in Sub-Saharan Africa, 1990–2017: Results from the Global Burden of Disease Study 2017. *Lancet Glob. Health* **2019**, *7*, e1375–e1387. [CrossRef] [PubMed]

17. Grundlingh, N.; Zewotir, T.T.; Roberts, D.J.; Manda, S. Assessment of Prevalence and Risk Factors of Diabetes and Pre-Diabetes in South Africa. *J. Health Popul. Nutr.* **2022**, *41*, 7. [CrossRef] [PubMed]
18. Agyei-Mensah, S.; de-Graft Aikins, A. Epidemiological Transition and the Double Burden of Disease in Accra, Ghana. *J. Urban Health Bull. N. Y. Acad. Med.* **2010**, *87*, 879–897. [CrossRef] [PubMed]
19. Naz, O.; Ibrahim, M.; Mohiuddin, A.F.; Khan, A.A.; Samad, Z. Public Health Data Quality and Evidence Use in Developing Countries: A Call to Action. *Front. Public Health* **2023**, *11*, 1194499. [CrossRef] [PubMed]
20. Kinyondo, A.; Pelizzo, R. Poor Quality of Data in Africa: What Are the Issues? *Polit. Policy* **2018**, *46*, 851–877. [CrossRef]
21. O'Neil, S.; Taylor, S.; Sivasankaran, A. Data Equity to Advance Health and Health Equity in Low- and Middle-Income Countries: A Scoping Review. *Digit. Health* **2021**, *7*, 20552076211061922. [CrossRef] [PubMed]
22. Mhlanga, D.; Garidzirai, R. The Influence of Racial Differences in the Demand for Healthcare in South Africa: A Case of Public Healthcare. *Int. J. Environ. Res. Public. Health* **2020**, *17*, 5043. [CrossRef] [PubMed]
23. Ahmed, S.K. Sample Size for Saturation in Qualitative Research: Debates, Definitions, and Strategies. *J. Med. Surg. Public Health* **2025**, *5*, 100171. [CrossRef]
24. Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C.W.; Choudhary, A.; Agrawal, A.; Billinge, S.J.L.; et al. Recent Advances and Applications of Deep Learning Methods in Materials Science. *NPJ Comput. Mater.* **2022**, *8*, 1–26. [CrossRef]
25. Goyal, M.; Mahmoud, Q.H. A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics* **2024**, *13*, 3509. [CrossRef]
26. Rajotte, J.-F.; Bergen, R.; Buckeridge, D.L.; El Emam, K.; Ng, R.; Strome, E. Synthetic Data as an Enabler for Machine Learning Applications in Medicine. *iScience* **2022**, *25*, 105331. [CrossRef] [PubMed]
27. Pezoulas, V.C.; Zaridis, D.I.; Mylona, E.; Androutsos, C.; Apostolidis, K.; Tachos, N.S.; Fotiadis, D.I. Synthetic Data Generation Methods in Healthcare: A Review on Open-Source Tools and Methods. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 2892–2910. [CrossRef] [PubMed]
28. Shahriar, S.; Al Roken, N. How Can Generative Adversarial Networks Impact Computer Generated Art? Insights from Poetry to Melody Conversion. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100066. [CrossRef]
29. Bourou, S.; El Saer, A.; Velivassaki, T.-H.; Voulkidis, A.; Zahariadis, T. A Review of Tabular Data Synthesis Using GANs on an IDS Dataset. *Information* **2021**, *12*, 375. [CrossRef]
30. Gm, H.; Gourisaria, M.K.; Pandey, M.; Rautaray, S.S. A Comprehensive Survey and Analysis of Generative Models in Machine Learning. *Comput. Sci. Rev.* **2020**, *38*, 100285. [CrossRef]
31. Alexander, M. South Africa's Population. South Africa Gateway 2025. Available online: https://southafrica-info.com/author/mary1alexander/ (accessed on 24 June 2025).
32. Konkor, I.; Kuuire, V.Z. Epidemiologic Transition and the Double Burden of Disease in Ghana: What Do We Know at the Neighborhood Level? *PLoS ONE* **2023**, *18*, e0281639. [CrossRef] [PubMed]
33. WHO SAGE Data Catalog. Available online: https://apps.who.int/healthinfo/systems/surveydata/index.php/catalog/sage/?page=1&ps=15&repo=sage (accessed on 12 May 2025).
34. Yoon, J.; Jordon, J.; van der Schaar, M. GAIN: Missing Data Imputation Using Generative Adversarial Nets. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
35. Mhawi, M.Y.; Abdullah, H.N.; Sikora, A. The Influence of Hyperparameters on GANs Performance for Medical Image Transformation. In Proceedings of the 2024 1st International Conference on Emerging Technologies for Dependable Internet of Things (ICETI), Sanaa, Yemen, 25–26 November 2024; pp. 1–8.
36. Serafim Rodrigues, T.; Rogério Pinheiro, P. Hyperparameter Optimization in Generative Adversarial Networks (GANs) Using Gaussian AHP. *IEEE Access* **2025**, *13*, 770–788. [CrossRef]
37. Patterson, J.; Gibson, A. *Deep Learning: A Practitioner's Approach*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017; ISBN 978-1-4919-1425-0.
38. Foster, D. *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2019; ISBN 978-1-4920-4191-7.
39. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv14062661. [CrossRef]
40. Hancock, J.T.; Khoshgoftaar, T.M. Survey on Categorical Data for Neural Networks. *J. Big Data* **2020**, *7*, 28. [CrossRef]
41. Otsu, T.; Taniguchi, G. Kolmogorov–Smirnov Type Test for Generated Variables. *Econ. Lett.* **2020**, *195*, 109401. [CrossRef]
42. Li, J.; Guo, S.; Ma, R.; He, J.; Zhang, X.; Rui, D.; Ding, Y.; Li, Y.; Jian, L.; Cheng, J.; et al. Comparison of the Effects of Imputation Methods for Missing Data in Predictive Modelling of Cohort Study Datasets. *BMC Med. Res. Methodol.* **2024**, *24*, 41. [CrossRef] [PubMed]
43. Xu, L. Synthesizing Tabular Data Using Conditional GAN. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2017.

44. Cheon, M.J.; Dong, H.L.; Lee, D.H.; Park, J.W.; Choi, H.J.; Lee, J.S.; Lee, O. CTGAN vs. TGAN? Which One is More Suitable for Generating Synthetic EEG Data. *J. Theor. Appl. Inf. Technol.* **2021**, *99*, 2359–2372.
45. Parise, O.; Kronenberger, R.; Parise, G.; de Asmundis, C.; Gelsomino, S.; La Meir, M. CTGAN-Driven Synthetic Data Generation: A Multidisciplinary, Expert-Guided Approach (TIMA). *Comput. Methods Programs Biomed.* **2025**, *259*, 108523. [CrossRef] [PubMed]
46. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular Data Using Conditional GAN. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
47. Pathare, A.; Mangrulkar, R.; Suvarna, K.; Parekh, A.; Thakur, G.; Gawade, A. Comparison of Tabular Synthetic Data Generation Techniques Using Propensity and Cluster Log Metric. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100177. [CrossRef]
48. Masarotto, G.; Varin, C. Gaussian Copula Marginal Regression. *Electron. J. Stat.* **2012**, *6*, 1517–1549. [CrossRef]
49. Houssou, R.; Augustin, M.-C.; Rappos, E.; Bonvin, V.; Robert-Nicoud, S. Generation and Simulation of Synthetic Datasets with Copulas. *arXiv* **2022**, arXiv:2203.17250.
50. Razghandi, M.; Zhou, H.; Erol-Kantarci, M.; Turgut, D. Variational Autoencoder Generative Adversarial Network for Synthetic Data Generation in Smart Home. In Proceedings of the ICC 2022–IEEE International Conference on Communications, Seoul, Republic of Korea, 16–20 May 2022; pp. 4781–4786.
51. Patki, N. The Synthetic Data Vault: Generative Modeling for Relational Databases. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2016.
52. Montanez, A. SDV: An Open Source Library for Synthetic Data Generation. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2018.
53. Campbell, J.N.; Dais Ferreira, M.; Isenor, A.W. Generation of Vessel Track Characteristics Using a Conditional Generative Adversarial Network (CGAN). *Appl. Artif. Intell.* **2024**, *38*, 2360283. [CrossRef]
54. SDMetrics. Available online: https://docs.sdv.dev/sdmetrics (accessed on 12 May 2025).
55. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
56. van Daalen, K.R.; Zhang, D.; Kaptoge, S.; Paige, E.; Angelantonio, E.D.; Pennells, L. Risk Estimation for the Primary Prevention of Cardiovascular Disease: Considerations for Appropriate Risk Prediction Model Selection. *Lancet Glob. Health* **2024**, *12*, e1343–e1358. [CrossRef] [PubMed]
57. Lloyd-Jones, D.M.; Braun, L.T.; Ndumele, C.E.; Smith, S.C.; Sperling, L.S.; Virani, S.S.; Blumenthal, R.S. Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report from the American Heart Association and American College of Cardiology. *Circulation* **2019**, *139*, e1162–e1177. [CrossRef] [PubMed]
58. Dong, W.; Fong, D.Y.T.; Yoon, J.; Wan, E.Y.F.; Bedford, L.E.; Tang, E.H.M.; Lam, C.L.K. Generative Adversarial Networks for Imputing Missing Data for Big Data Clinical Research. *BMC Med. Res. Methodol.* **2021**, *21*, 78. [CrossRef] [PubMed]