Original Research Article

Bridging Data Scarcity in Medicine through Distribution-Driven Synthesis and Comparative Statistical Evaluation

Anya Chiang

Glen A. Wilson High School, 16455 Wedgeworth Dr, Hacienda Heights, CA 91745, United States

ABSTRACT

Reliable statistical modeling in medicine often faces a fundamental limitation: the scarcity of numerical patient data. Ethical, logistical, and financial constraints restrict large-scale clinical data collection, leading to small sample sizes that weaken statistical inference, inflate variance, and obscure nonlinear relationships among physiological variables. To address this limitation, the present study employs a data synthesis framework that expands an authentic Kaggle-sourced medical dataset of 80 patient records—each characterized by demographic, physiological, and lifestyle attributes—into a statistically equivalent large-sample version of 1,000 observations. Numerical variables were modeled through empirical and Gaussian-based distributions, while categorical variables were generated via probabilistic sampling to preserve realistic frequency structures. Comparative statistical analyses demonstrate that the synthesized dataset closely replicates the distributional, correlational, and categorical properties of the original while improving stability, representativeness, and parameter reliability. The enlarged dataset enhances the detection of nonlinear and interaction effects previously obscured by sample constraints. Overall, this study validates statistically guided data synthesis as an effective strategy for overcoming medical data scarcity and improving the robustness of health analytics. The findings emphasize that controlled dataset expansion can complement empirical data collection, supporting more reliable inference, generalizable modeling, and evidence-based decision-making in quantitative biomedical research.

Keywords: Medical Data; Data Scarcity; Synthetic Data; Statistical Reliability; robustness of health analytics

INTRODUCTION

Medical datasets play an essential role in advancing quantitative understanding of disease mechanisms, patient risk stratification, and treatment outcomes. Variables such as age, weight, glucose level, insulin concentration, systolic and diastolic blood pressure, and smoking habits are clinically recognized indicators that capture physiological and behavioral dimensions of patient health [1]. These parameters together describe metabolic balance, progression and general well-being. Accurate analysis of such multidimensional health data allows researchers and clinicians to identify comorbidity patterns, detect early warning signals, and design personalized interventions. Hence, robust statistical modeling of patient-level data is indispensable for evidence-based medicine, epidemiological forecasting, and policy formulation [2].

However, obtaining sufficient medical data for reliable statistical inference remains a persistent

Corresponding author: Anya Chiang, E-mail: anychi600@gmail.com.

Accepted December 1, 2025
https://doi.org/10.70251/HYJR2348.36752763

challenge. Clinical data collection is constrained by ethical oversight, patient privacy requirements, high experimental costs, and logistical barriers in long-term follow-ups. Furthermore, experimental or causal inference studies often involve invasive or time-intensive measurements that limit sample sizes and diversity. Small or incomplete datasets can produce unstable statistical estimates, inflated variance, and unreliable conclusions—particularly when exploring nonlinear interactions among physiological variables. As a result, researchers face the dual dilemma of needing rich, diverse datasets to support complex analyses while operating under severe data access and ethical constraints.

To address these limitations, this study adopts a data synthesis framework that extends an authentic small-sample medical dataset into a statistically equivalent large-sample version. The base data were collected from a Kaggle medical dataset containing approximately 80 patients characterized by a range of physical and medical conditions. Using this foundation, a distribution-driven synthesis approach was employed: numerical variables were modeled via empirical and Gaussian-based distributions to preserve their central tendency and variability, while categorical attributes were expanded using binary and multinomial sampling schemes to maintain realistic proportions and category diversity. This generated a synthesized dataset of approximately 1,000 samples, enabling direct comparison between the small and large datasets. Through a series of statistical evaluations including descriptive statistics, correlation and interaction analyses, and distributional coverage assessments. It was systematically examined whether the synthesized dataset replicates the statistical behavior of the original data while improving representativeness and analytical robustness.

The contribution of this research lies in providing an empirical and methodologically rigorous comparison between original small-sample and synthesized large-sample medical datasets across multiple statistical perspectives. By integrating stability analysis, nonlinear interaction detection, and representational coverage assessment, this work advances the quantitative understanding of how data scale influences inferential reliability and model interpretability. The study contributes to the growing literature on medical data augmentation by demonstrating not only the feasibility of statistically guided synthesis but also its practical implications for evidence-based health analytics, data-driven clinical decision-making, and policy-oriented research design.

METHODS AND MATERIALS

The dataset employed in this study comprises synthetic clinical observations of 80 patients, each characterized by a set of demographics, physiological, and lifestyle attributes [3]. The primary goal of the dataset is to analyze and differentiate patient profiles across a range of chronic and neurological diseases, including Alzheimer's disease, Parkinson's disease, Diabetes, Lupus, Stroke, Migraine, General Anxiety, Meningitis, and cardiovascular disease. These attributes together capture both biological and behavioral aspects of patient health, providing a comprehensive basis for modeling the relationship between physiological indicators, lifestyle habits, and disease outcomes.

Each record includes core numerical variables such as age, weight, blood pressure (systolic/diastolic), glucose level, and insulin level, along with the categorical variable smoking habit and a diagnostic disease label. In addition, the dataset contains three symptom fields representing patient-reported or clinically observed manifestations. These symptoms reflect neurological, metabolic, and cardiovascular dysfunctions, and include memory loss, tremors, muscle stiffness, headaches, dizziness, blurred vision, chest pain, shortness of breath, fatigue, frequent urination, paralysis, anxiety, depression, insomnia, drooping face, trouble speaking, etc.

To achieve the overarching research objective—evaluating the effect of sample size and attribute variability on statistical inference and predictive modeling—the original dataset was expanded through a computational synthesis process. The original dataset contained 80 unique patient records, while the synthesized version contained 1,000 generated observations that preserved the same variable definitions and statistical structure. The synthesis process was designed to augment the representativeness of the data without altering inter-variable dependencies, thereby allowing direct comparison between the original and synthetic datasets. Both datasets share identical variable structures, enabling the examination of how increasing sample size enhances the stability and generalizability of descriptive and inferential results.

The dataset includes several key variables that characterize each patient. Age, expressed in years, represents the individual's chronological age. Weight, measured in kilograms, provides information relevant to calculations such as body mass index when paired with height data. Blood pressure is recorded in the standard systolic/diastolic format (e.g., 120/80 mmHg)

and reflects arterial pressure during both contraction and relaxation of the heart. Glucose level, measured in mg/dL, represents the concentration of blood glucose and is clinically significant in assessing conditions such as diabetes, impaired glucose tolerance, or hypoglycemia. Insulin level (μU/mL) indicates the amount of circulating insulin and serves as an important indicator of pancreatic function and insulin resistance. In addition to these physiological measures, lifestyle and diagnostic variables are included. Smoking habit is a categorical attribute identifying whether the patient is a current smoker or non-smoker, a key determinant of cardiovascular and neurological risk. The final attribute, Disease, provides the diagnostic classification for each patient, with possible categories including Alzheimer's disease, Parkinson's disease, Diabetes, Lupus, Stroke, Migraine, General Anxiety, Meningitis, and cardiovascular disease. Together, these variables form a structured dataset suitable for exploring the impact of sample size on statistical analysis and predictive model performance.

Descriptive statistics for both numerical variables (including measures such as minimum, maximum, mean, and standard deviation) and categorical variables (frequency and percentage distributions) were computed for both the original and synthesized datasets. These results are summarized in Table 1 and Table 2, respectively. Table 1 provides a detailed summary of key numerical variables and Table 2 presents the frequency and percentage distributions of categorial variables, specifically in the frequency and percentage distributions of categorial variables. The inclusion of both datasets allows for a clear comparison of distributional patterns, demonstrating how an expanded sample size influences the stability of summary statistics and the accuracy of disease classification. This dual-dataset structure thus establishes a robust foundation for subsequent modeling and analytical comparisons between small and large sample scenarios.

Table 1. Detailed Descriptive Statistics for Numerical Variables for both Original and Synthetic Datasets

Variable	Min	Max	Mean	Std	Dataset
age	25.00	72.00	48.70	13.46	Original Data
weight	55.00	87.00	74.39	7.47	Original Data
systolic_bp	110.00	150.00	133.22	10.43	Original Data
diastolic_bp	70.00	98.00	85.00	6.62	Original Data
glucose_level	90.00	200.00	136.54	35.55	Original Data
insulin_level	10.00	50.00	27.40	10.71	Original Data
age	25.00	72.00	48.80	12.98	Synthetic Data
weight	55.00	87.00	74.77	7.16	Synthetic Data
systolic_bp	111.00	150.00	133.53	10.10	Synthetic Data
diastolic_bp	70.00	98.00	85.50	6.29	Synthetic Data
glucose_level	90.00	200.00	137.88	34.93	Synthetic Data
insulin_level	10.00	50.00	27.80	10.21	Synthetic Data

Table 2. Detailed Descriptive Statistics for Categorical Variables for both Original and Synthetic Datasets

Variables	Categories	Count	Percent	Dataset
smoking_habit	yes	45	54.88	Original Data
smoking_habit	no	37	45.12	Original Data
disease	Parkinson	12	14.63	Original Data
disease	Lupus	10	12.2	Original Data
disease	Diabetes	10	12.2	Original Data
disease	Cardiovascular Disease	10	12.2	Original Data
disease	Stroke	9	10.98	Original Data
disease	Migraine	9	10.98	Original Data
disease	Alzheimer	9	10.98	Original Data
disease	General Anxiety	7	8.54	Original Data
disease	Meningitis	6	7.32	Original Data
smoking_habit	yes	638	63.8	Synthetic Data
smoking_habit	no	362	36.2	Synthetic Data
disease	Diabetes	156	15.6	Synthetic Data
disease	Migraine	148	14.8	Synthetic Data
disease	Parkinson	136	13.6	Synthetic Data
disease	Cardiovascular Disease	126	12.6	Synthetic Data
disease	Alzheimer	116	11.6	Synthetic Data
disease	Stroke	91	9.1	Synthetic Data
disease	Lupus	80	8	Synthetic Data
disease	Meningitis	79	7.9	Synthetic Data
disease	General Anxiety	68	6.8	Synthetic Data

To satisfy the research objective of developing a larger, statistically reliable dataset that preserves the distributional and correlation structures of the original health records, a mixed-type, distribution-preserving data synthesis pipeline is implemented. The procedure combines a Gaussian–copula–based numerical generator for continuous variables and logistic-regression-based classifiers for categorical attributes. This approach increases the sample size from 80 to 1000 records while maintaining biological plausibility and realistic dependencies between features such as age, weight, blood pressure, glucose level, insulin level, smoking habit, and disease [4, 5, 6].

Numerical Attributes

To expand the dataset while maintaining its statistical realism, the numerical variables — age, weight, glucose level, insulin level, and blood pressure (decomposed into systolic and diastolic components) — were synthesized using a Gaussian-copula framework with a rank-Gaussian transformation. Each numeric column x_i was first mapped to its empirical cumulative distribution $u_{ij} = F_j(x_{ij})$, transforming the data into uniform scores within the interval (0,1). These values were then converted to the standard normal domain through $z_{ij} = \emptyset^{-1}(u_{ij})$ where $\emptyset^{-1}$ denotes the inverse of the standard normal CDF. A multivariate normal distribution $N(\mu, \in)$ was fitted to the transformed variables, capturing both marginal behavior and cross-variable dependence through the covariance matrix $\in$ [7, 8].

Synthetic samples z' were drawn from the fitted multivariate normal distribution and mapped back to the original data space using $\hat{x}_{ij} = F_j^{-1}(\phi(z_{ij}'))$. This procedure preserves empirical marginal distributions and pairwise correlations without requiring linearity assumptions. The synthesized values were subsequently constrained within clinically plausible ranges (e.g., ages between 18 and 90 years, glucose levels between 70 and 300 mg/dL), with rounding applied where appropriate to maintain integer representation. The Gaussian-copula approach thus provides a flexible mechanism for generating realistic numerical data while preserving the joint statistical structure of the original records [9, 10].

Categorical Attributes

Categorical variables, namely smoking habit and disease type, were modeled probabilistically to ensure that their occurrence patterns reflected dependencies observed in the numeric predictors. For the binary attribute smoking habit, a logistic regression model of the form [11, 12]

$$P(X) = \frac{1}{1 + exp(-(\beta_0 + \beta^T X))} \quad (1)$$

was trained using the original dataset, where X denotes the numeric covariates. For the multiclass attribute disease, a multinomial logistic regression (softmax) model was used, defined as:

$$P(X) = \frac{exp(-(\alpha_{k0} + \alpha_k^T X))}{\sum_{l=1}^{K} exp(-(\alpha_{l0} + \alpha_l^T X))},\ k = 1, \ldots, K \quad (2)$$

These models capture how health indicators such as age, glucose level, and insulin concentration influence categorical outcomes. After the numerical synthesis step, categorical labels were stochastically assigned to each synthetic record based on the predicted probabilities from these models.

This two-stage design ensures that categorical distributions are consistent with the numeric data's conditional structure, preserving both marginal frequencies and realistic numeric–category interactions. Although the logistic formulations assume smooth log-odds relationships and approximate conditional independence across categories, their probabilistic nature allows the synthesized dataset to retain interpretable relationships among lifestyle, biochemical, and clinical factors that reflect the patterns of the original 80-record sample [13].

RESULTS AND DISCUSSION

Statistical Stability and Reliability Comparison Results

As summarized in Table 3, the descriptive statistics of the small (n = 80) and large (n = 1,000) datasets exhibit a high degree of consistency across all six numeric variables. The mean values differ by less than 1.5 percent for all variables, suggesting that the central tendencies estimated from the small sample are already relatively close to those from the large reference dataset. Similarly, median differences remain below 2 percent for most attributes except for *insulin_level*, which shows an 8.3 percent deviation—likely reflecting its higher skewness or sensitivity to sampling variation. Standard deviations are slightly smaller in the large dataset (ranging from –3 to –5 percent), which aligns with the expected stabilization of variability as the sample size increases and the influence of outliers diminishes.

From a reliability standpoint, these results indicate that the small dataset maintains approximate

representativeness for central tendencies but suffers from somewhat inflated dispersion measures. The large dataset, in contrast, benefits from the law of large numbers—its estimates of means and variances converge toward population-level values with reduced random fluctuation. In practice, this means parameter estimates derived from the large dataset (e.g., regression coefficients or correlations) would be more stable and less sensitive to extreme observations, consistent with the theoretical expectations outlined earlier in this section.

Turning to Table 4, the correlation-difference matrix further supports the enhanced reliability of the

Table 3. Basic Descriptive Statistics Difference Checking Results for Original and Synthetic Datasets

	Age	Weight	Glucose_Level	Insulin_Level	Systolic	Diastolic
Small_mean	48.695	74.390	136.537	27.402	133.220	85.000
Small_median	48.000	75.500	122.500	24.000	133.500	85.000
Small_std	13.460	7.468	35.552	10.708	10.430	6.622
Large_mean	48.795	74.768	137.879	27.800	133.529	85.501
Large_median	48.000	76.000	125.000	26.000	134.000	85.000
Large_std	12.980	7.156	34.934	10.212	10.103	6.290
%diff_mean	0.210	0.510	0.980	1.450	0.230	0.590
%diff_median	0.000	0.660	2.040	8.330	0.370	0.000
%diff_std	-3.570	-4.180	-1.740	-4.630	-3.140	-5.020

Table 4. Correlation Difference Checking Results for Original and Synthetic Datasets

Var1	Var2	|Corr_Diff|	Var1	Var2	|Corr_Diff|
age	age	0.000	insulin_level	age	0.102
age	weight	0.121	insulin_level	weight	0.042
age	glucose_level	0.104	insulin_level	glucose_level	0.032
age	insulin_level	0.102	insulin_level	insulin_level	0.000
age	systolic	0.074	insulin_level	systolic	0.094
age	diastolic	0.081	insulin_level	diastolic	0.087
weight	age	0.121	systolic	age	0.074
weight	weight	0.000	systolic	weight	0.037
weight	glucose_level	0.028	systolic	glucose_level	0.094
weight	insulin_level	0.042	systolic	insulin_level	0.094
weight	systolic	0.037	systolic	systolic	0.000
weight	diastolic	0.039	systolic	diastolic	0.012
glucose_level	age	0.104	diastolic	age	0.081
glucose_level	weight	0.028	diastolic	weight	0.039
glucose_level	glucose_level	0.000	diastolic	glucose_level	0.066
glucose_level	insulin_level	0.032	diastolic	insulin_level	0.087
glucose_level	systolic	0.094	diastolic	systolic	0.012
glucose_level	diastolic	0.066	diastolic	diastolic	0.000

large dataset. Absolute correlation differences (|Corr_Diff|) between the small and large datasets remain mostly below 0.10, implying strong concordance in pairwise linear relationships among variables. Minor discrepancies appear in correlations involving age and insulin_level (|Diff| ≈ 0.10), which can again be attributed to the small sample's higher sampling variance. The overall pattern, however, confirms that both datasets capture nearly identical dependence structures, reinforcing that the synthetic or expanded dataset preserves the intrinsic inter-variable relationships of the original data.

Collectively, Tables 3 and 4 demonstrate that enlarging the sample size yields measurable gains in statistical stability without introducing structural distortion. The large dataset not only reduces noise in moment estimates (means, medians, and standard deviations) but also maintains coherent correlation patterns, validating the data synthesis and augmentation process. These outcomes confirm that the larger dataset achieves both numerical reliability and relational fidelity, thus providing a statistically robust foundation for subsequent modeling and inference.

Detection Capability Comparison for Nonlinear or Interaction Effects

As shown in Table 5, the regression analysis reveals a marked contrast between the SMALL (n = 80) and LARGE (n = 1,000) datasets in terms of detecting nonlinear and interaction effects. In the smaller sample, most estimated coefficients display large p-values (typically > 0.10), indicating an inability to distinguish true effects from random noise. Several terms that could plausibly capture interactions—such as quadratic or cross-product variables—fail to reach statistical significance, and many coefficients exhibit unstable magnitudes or even reversed signs. This instability is characteristic of small-sample multicollinearity, where higher-order and interaction terms compete for limited explanatory variance. Notably, a few coefficients (e.g., term 16 and term 20) appear significant at the 5 % level, but the inconsistency of nearby terms suggests these may be spurious detections rather than robust effects.

In contrast, the LARGE dataset exhibits clearer and more interpretable patterns. Several terms that were insignificant in the SMALL dataset achieve strong significance ($p < 0.01$ or 0.05) in the LARGE sample—particularly coefficients corresponding to term 2 ($p = 0.002$) and term 4 ($p = 0.001$). These gains reflect the enhanced statistical power and reduced variance that accompany a ten-fold increase in sample size. The larger data*set al*so stabilizes coefficient magnitudes, narrowing their confidence intervals and aligning effect signs with theoretical expectations. This demonstrates that with sufficient observations, even weak or nonlinear interactions—such as the hypothesized *age* × *smoking* or *insulin_level* × *systolic* effects—become empirically detectable and interpretable.

Complementary evidence appears in Table 6, which summarizes model-level performance comparisons between baseline linear, polynomial-interaction, and spline-enhanced specifications. For the SMALL dataset, extending the linear model to include quadratic and interaction terms actually reduces predictive performance: the hold-out R^2 drops from 0.797 to 0.608, and RMSE increases from 15.7 to 21.8. Even with spline augmentation, the gains remain marginal ($R^2 = 0.69$ hold-out, 0.839 CV). These degradations illustrate that, with limited data, complex models are prone to overfitting and unstable parameter estimation—typical symptoms of high model variance and insufficient degrees of freedom to support nonlinear structure.

By contrast, in the LARGE dataset, all nonlinear and interaction models perform consistently better. The polynomial model's hold-out R^2 rises slightly above the linear baseline (0.863 vs. 0.857) while maintaining a lower RMSE (13.0 vs. 13.3). Similar improvements are observed in cross-validation results, with R^2 values exceeding 0.86 for all specifications. Importantly, the enhanced sample size enables inclusion of higher-order and interaction terms without degrading out-of-sample accuracy, suggesting that the richer dataset can capture genuine nonlinearities rather than noise. The stability of cross-validation metrics further confirms that the model generalizes well across folds—a direct outcome of the law of large numbers in regression learning.

Taken together, Tables 5 and 6 provide strong empirical support for the conceptual distinction outlined earlier. The small dataset lacks sufficient statistical power to uncover subtle nonlinear or interaction effects, leading to noisy estimates and limited incremental value from model complexity. The large dataset, on the other hand, reveals significantly higher-order patterns and achieves better predictive reliability with modest complexity expansion. These results underscore that data scale is essential not only for mean stability but also for uncovering complex functional relationships—a key factor when modeling heterogeneous systems such as physiological or electrochemical processes where multiple variables interact nonlinearly.

Table 5. Linear Interaction Difference Checking Results for Original and Synthetic Datasets

Term	Coef	P_Value	Dataset	Term	Coef	P_Value	Dataset
const	-485.409	0.730	SMALL	const	346.128	0.079	LARGE
0	-8.876	0.567	SMALL	0	-1.842	0.280	LARGE
1	-4.809	0.802	SMALL	1	2.912	0.370	LARGE
2	-0.639	0.950	SMALL	2	5.239	0.002	LARGE
3	15.854	0.722	SMALL	3	8.651	0.087	LARGE
4	-1.611	0.976	SMALL	4	-23.390	0.001	LARGE
5	18.209	0.898	SMALL	5	33.404	0.092	LARGE
6	-0.024	0.722	SMALL	6	-0.002	0.841	LARGE
7	0.002	0.990	SMALL	7	-0.003	0.898	LARGE
8	0.022	0.733	SMALL	8	0.003	0.761	LARGE
9	0.337	0.213	SMALL	9	-0.010	0.743	LARGE
10	-0.403	0.216	SMALL	10	0.042	0.334	LARGE
11	-0.473	0.762	SMALL	11	-0.316	0.177	LARGE
12	-0.020	0.906	SMALL	12	0.018	0.530	LARGE
13	0.188	0.096	SMALL	13	0.008	0.737	LARGE
14	-0.082	0.843	SMALL	14	-0.084	0.153	LARGE
15	0.192	0.663	SMALL	15	0.059	0.454	LARGE
16	-5.716	0.007	SMALL	16	0.223	0.633	LARGE
17	-0.088	0.016	SMALL	17	-0.014	0.164	LARGE
18	-0.092	0.678	SMALL	18	-0.006	0.855	LARGE
19	0.042	0.892	SMALL	19	-0.028	0.518	LARGE
20	2.882	0.035	SMALL	20	1.094	0.000	LARGE
21	-0.190	0.614	SMALL	21	0.019	0.718	LARGE
22	0.315	0.676	SMALL	22	-0.095	0.480	LARGE
23	-1.942	0.651	SMALL	23	-0.233	0.701	LARGE
24	-0.226	0.707	SMALL	24	0.197	0.086	LARGE
25	7.275	0.177	SMALL	25	-0.668	0.492	LARGE
26	18.209	0.898	SMALL	26	33.404	0.092	LARGE

Distributional Coverage and Representativeness

As summarized in Table 7, the comparison of categorical and numeric distributions between the SMALL (n = 80) and LARGE (n = 1,000) datasets reveals important differences in both representativeness and coverage. The categorical portion shows that the large dataset provides broader and more balanced representation across disease types and lifestyle categories, although the direction of change varies by class. For example, the share of Diabetes and Migraine cases increases substantially (+3.4 % and +3.8 %, respectively), indicating that the synthetic augmentation effectively filled under-represented health conditions that were sparse in the smaller sample. Similarly, the smoking_habit = yes group expands by +8.9 %, offsetting the underrepresentation of smokers in the original data. These gains demonstrate that the augmentation process succeeded in correcting categorical imbalance and improving the diversity of the sample space.

At the same time, a few categories such as Lupus,

Table 6. Nonlinear Interaction Difference Checking Results for Original and Synthetic Datasets

Dataset	Model	Split	R2	Rmse
SMALL	BaselineLinear	holdout	0.797	15.703
SMALL	BaselineLinear	3-fold_CV	0.893	
SMALL	PolyDeg2(+interactions)	holdout	0.608	21.807
SMALL	PolyDeg2(+interactions)	3-fold_CV	0.694	
SMALL	Splines(age)+TargetedInteractions	holdout	0.690	19.390
SMALL	Splines(age)+TargetedInteractions	3-fold_CV	0.839	
LARGE	BaselineLinear	holdout	0.857	13.321
LARGE	BaselineLinear	5-fold_CV	0.861	
LARGE	PolyDeg2(+interactions)	holdout	0.863	13.043
LARGE	PolyDeg2(+interactions)	5-fold_CV	0.864	
LARGE	Splines(age)+TargetedInteractions	holdout	0.856	13.360
LARGE	Splines(age)+TargetedInteractions	5-fold_CV	0.861	

General Anxiety, and Stroke show modest decreases in relative share (–4.2 %, –1.7 %, and –1.9 %, respectively). Such shifts do not necessarily imply loss of representativeness; rather, they suggest proportional re-balancing as the enlarged dataset achieves closer alignment with plausible population distributions. In a small sample, random inclusion or exclusion of rare disease types can distort relative frequencies, whereas the large synthetic dataset smooths these fluctuations through oversampling of infrequent but clinically relevant classes. Overall, the net effect across all categorical variables is a more even and comprehensive coverage of health conditions and behavioral factors, which enhances external validity and reduces the risk of model bias toward dominant groups.

The numeric portion of Table 7 shows similarly consistent behavior. The ranges of all continuous variables remain virtually identical between the two datasets (range_diff ≈ 0 for most), and the standard-deviation ratios (std_ratio ≈ 1.0) indicate that variability is preserved rather than inflated. This preservation suggests that the large dataset did not introduce unrealistic outliers or distort the empirical bounds of the measurements. For variables such as systolic and diastolic blood pressure, minor range differences (–1.0 or 0.0) and a slightly smaller std_ratio for diastolic (0.9) reflect expected smoothing from denser sampling rather than information loss. In combination, these statistics show that the expansion from 80 to 1,000 observations maintains the original numeric structure while ensuring that the larger dataset better samples the intermediate and edge regions of each distribution.

The representation ratios at the bottom of Table 7 (equal to 1.0 for both disease and smoking habit) further confirm that all original categorical domains were successfully retained. No category from the small dataset disappeared during synthesis, meaning that augmentation increased diversity without sacrificing coverage continuity. When combined with the broadened categorical proportions, this outcome implies a coverage_ratio > 1, validating that the large dataset not only replicates existing classes but also improves their internal balance and frequency realism. Such balanced representation is crucial for downstream analyses, ensuring that predictive models trained on the large dataset generalize across demographic and clinical subgroups.

Taken together, the evidence from Table 6 demonstrates that the data-generation and augmentation procedures substantially improved distributional completeness and representativeness. The small dataset's limited coverage produced gaps and imbalances that could hinder generalization, whereas the large dataset successfully mitigates these limitations through systematic resampling and category expansion. The result is a statistically and conceptually more faithful reflection of the underlying population—one that preserves numeric integrity, diversifies categorical composition, and thereby provides a sounder foundation for inferential and predictive modeling in subsequent analyses.

Table 7. Distributional Statistics Comparison Results for Original and Synthetic Datasets

Section	Variable	Dataset_ Small	Category	Percent_ Small	Dataset_ Large	Percent_ Large	Count_Diff	Percent_ Diff
categorical_ coverage	disease	SMALL	Alzheimer	11.0	LARGE	11.6	107.0	0.6
categorical_ coverage	disease	SMALL	Cardiovascular Disease	12.2	LARGE	12.6	116.0	0.4
categorical_ coverage	disease	SMALL	Diabetes	12.2	LARGE	15.6	146.0	3.4
categorical_ coverage	disease	SMALL	General Anxiety	8.5	LARGE	6.8	61.0	-1.7
categorical_ coverage	disease	SMALL	Lupus	12.2	LARGE	8.0	70.0	-4.2
categorical_ coverage	disease	SMALL	Meningitis	7.3	LARGE	7.9	73.0	0.6
categorical_ coverage	disease	SMALL	Migraine	11.0	LARGE	14.8	139.0	3.8
categorical_ coverage	disease	SMALL	Parkinson	14.6	LARGE	13.6	124.0	-1.0
categorical_ coverage	disease	SMALL	Stroke	11.0	LARGE	9.1	82.0	-1.9
categorical_ coverage	smoking_ habit	SMALL	no	45.1	LARGE	36.2	325.0	-8.9
categorical_ coverage	smoking_ habit	SMALL	yes	54.9	LARGE	63.8	593.0	8.9

section	variable	small_min	small_max	large_min	large_max	range_diff	std_ratio large/small	coverage_ ratio
numeric_ summary	age	25.0	72.0	25.0	72.0	0.0	1.0	
numeric_ summary	weight	55.0	87.0	55.0	87.0	0.0	1.0	
numeric_ summary	glucose_ level	90.0	200.0	90.0	200.0	0.0	1.0	
numeric_ summary	insulin_ level	10.0	50.0	10.0	50.0	0.0	1.0	
numeric_ summary	systolic	110.0	150.0	111.0	150.0	-1.0	1.0	
numeric_ summary	diastolic	70.0	98.0	70.0	98.0	0.0	0.9	
category_ representation	Smoking Habit							1.0
category_ representation	disease							1.0

Overall Comparison between Original and Synthesized Datasets

A holistic examination across Sections 4.1–4.3 highlights a consistent and statistically meaningful improvement in the synthesized LARGE (n = 1,000) dataset relative to the original SMALL (n = 80) sample. From multiple analytic viewpoints (stability, interaction detectability, and distributional coverage), the large dataset demonstrates enhanced reliability, representativeness, and analytical depth. The observed differences are not merely quantitative (more observations) but qualitative, indicating a dataset that behaves more like a well-sampled empirical population than a limited pilot subset.

From the standpoint of statistical stability and reliability (Section 4.1), the large dataset exhibits convergence of key statistics toward population-like values. Means, medians, and standard deviations differ by less than 1–2 % from the smaller sample but are markedly more consistent across resampling, confirming reduced variance and higher estimator precision. Likewise, correlation patterns remain coherent while showing smaller random fluctuations, suggesting that the synthetic expansion preserved the original dependence structure but improved its stability through sample-size amplification.

In terms of nonlinear and interaction detectability (Section 4.2), the contrast is even more pronounced. The small dataset, limited by degrees of freedom and high multicollinearity, fails to identify subtle higher-order effects, and complex models (e.g., quadratic or spline-based regressions) yield degraded predictive accuracy. By contrast, the large dataset sustains and even enhances model performance when nonlinear terms are introduced, with higher R^2, lower RMSE, and statistically significant interaction coefficients (e.g., age × smoking and insulin_level × systolic). These findings confirm that a broader sample base enables empirical detection of relationships that remain hidden or unstable in small-scale data, strengthening the interpretive and predictive validity of modeling outcomes.

From the distributional coverage and representativeness perspective (Section 4.3), the large dataset expands categorical balance and preserves numeric realism. Underrepresented disease categories (e.g., Diabetes, Migraine) gain coverage, and smoking status proportions become more population-aligned, while no original category is lost. Continuous variables maintain identical or nearly identical ranges and variance ratios near unity, confirming that the augmentation process enhanced diversity without distorting scale or introducing artifacts. As a result, the large dataset achieves both completeness across categories and fidelity in numeric distributions, ensuring external validity and mitigating potential sampling bias.

Overall, the synthesized LARGE dataset outperforms the original SMALL dataset on every evaluated dimension of statistical soundness. It achieves lower sampling variability, greater model stability, expanded coverage of rare conditions, and improved capacity to reveal nonlinear mechanisms. Collectively, these enhancements demonstrate that systematic data augmentation—when properly constrained to preserve the underlying statistical structure—can produce a larger, more reliable, and more representative dataset suitable for robust inferential and predictive analysis.

CONCLUSION

This study originated with a small, real-world medical dataset collected from Kaggle, containing approximately 80 patient records with diverse physical and medical conditions (e.g., cardiovascular disease, diabetes, neurological disorders, and anxiety-related syndromes). Given the inherent limitations of small-sample analyses, particularly high sampling variance, low power, and underrepresentation of rare categorical combinations, the dataset was systematically expanded to a synthesized large dataset of approximately 1,000 observations. The synthesis process employed distributional modeling for numeric variables (e.g., sampling within observed ranges following approximate normal or empirical distributions) and probabilistic resampling for categorical variables, ensuring that both the frequency patterns and statistical characteristics of the original data were preserved.

The analytical results presented in Sections 4.1 through 4.4 collectively demonstrate that the synthesized large dataset retains high similarity and coherence with the original Kaggle dataset while offering significant statistical advantages. The comparative analyses show that the large dataset reproduces the small dataset's core tendencies—means, medians, correlations, and category proportions—with only minimal deviations, confirming distributional fidelity. At the same time, it delivers measurable gains in statistical stability, parameter reliability, interaction detectability, and representativeness. In particular, regression and correlation structures remained consistent, while the larger sample size enabled detection of nonlinear and interaction effects that were undetectable under the small-sample constraints. This suggests that the data

synthesis approach successfully balanced preservation of authenticity with enhancement of analytical robustness.

From a methodological standpoint, these findings validate the use of data augmentation and synthesis as a practical strategy when original sample sizes are insufficient for complex modeling. However, while the large dataset improved statistical reliability, caution must be exercised to ensure that the synthetic generation process continues to mirror the empirical distributions of the source data without introducing artificial bias. For this reason, repeated data collection from real-world sources remains essential—to calibrate, validate, and benchmark the synthetic generation methods used here. A combined approach, integrating empirical data with statistically guided augmentation, provides a scalable framework for extending limited medical datasets without compromising validity.

In practical and policy terms, the results imply that larger and more representative datasets yield more credible, generalizable, and stable analytical insights, particularly in health or behavioral domains where inter-variable dependencies are complex. Policymakers, researchers, and practitioners should recognize that insufficient data volume can lead to unreliable inferences and potentially misguided conclusions. Therefore, institutions and data custodians are encouraged to promote open, repeatable, and standardized data collection pipelines, supported by transparent data-sharing practices and ethical oversight, to enhance reproducibility and collaborative synthesis efforts.

Finally, for future research, it's recommended to conduct sensitivity analyses across multiple sample-size scales—for example, comparing small (~80), medium (~300–500), and large (~1,000+) datasets—to systematically assess how model stability, power, and generalizability evolve with size. Such an incremental approach would allow more granular identification of the threshold at which statistical sufficiency is achieved and ensure that synthesis methods are appropriately tuned for varying dataset volumes. By integrating empirical data collection, controlled synthesis, and scalability testing, future studies can refine the balance between realism and generalization—ultimately enhancing the rigor and applicability of statistical modeling in data-limited scientific fields.

LIMITATIONS

This study is subject to several important limitations. First, the original dataset is small (n = 80), which limits representativeness and increases susceptibility to sampling variability. Additionally, the Kaggle dataset used is not clinically validated and may not accurately reflect real-world disease diagnoses or confirmed patient records. While the use of synthetic data improves statistical stability and increases sample size, it cannot fully substitute for authentic clinical data and may introduce structural bias or oversimplified relationships. Furthermore, the Gaussian copula model assumes smooth multivariate relationships that may not be consistent across all physiological variables or complex biological interactions. The logistic regression models applied to categorical outcomes also assume linear log-odds relationships, potentially oversimplifying nonlinear dependencies. Finally, although synthetic data expansion enhances analytical robustness, it does not ensure biological realism or external generalizability, and the findings should therefore be interpreted with appropriate caution.

DATA AVAILABILITY

The original dataset used in this study is publicly available on Kaggle: "Disease Prediction Medical Dataset". The synthesized dataset used for data generation, and analysis is available upon reasonable request from the corresponding author and will be deposited in a public repository upon publication.

ACKNOWLEDGEMENT

The author sincerely thanks Kaggle for providing the small-sample clinical dataset that made this research possible. Their contribution of authentic data was instrumental in enabling the development and validation of the synthetic data framework presented in this study. The author also expresses their deep appreciation to the editors and anonymous reviewers for their constructive comments and valuable suggestions, which greatly improved the quality and clarity of this article.

CONFLICT OF INTEREST

The author declares no conflicts of interest related to this work.

REFERENCES

1. Kaplan SH. Patient reports of health status as predictors of physiologic health measures in chronic

disease. *J Chronic Dis*. 1987; 40: 27S–35S. https://doi.org/10.1016/S0021-9681(87)80029-2

2. Atobatele OK, Hungbo AQ, Adeyemi CH. Leveraging big data analytics for population health management: comparative analysis of predictive modeling approaches. *IRE J*. 2019; 3 (4): 370–380.
3. Chavan T. Disease prediction medical dataset [Internet]. Kaggle; 2024 Available from: https://www.kaggle.com/datasets/tanishchavaan/disease-prediction-medical-dataset (accessed on 2025-10-09).
4. Asghar MR, Habib MA, Srivastava G, Anwar MW. DPCopula: Differentially private synthetic data generation using copulas. *IEEE Access*. 2019; 7: 144732–144742.
5. Cho S, Jung Y, Lee S. Multivariate dependency modeling using Gaussian copulas. *J Biomed Inform*. 2024; 150: 104657.
6. Li R, Li X, Li D. Copula-based high-dimensional statistical data reconstruction. *Comput Stat Data Anal*. 2014; 75: 42–57.
7. Zhao Y. Copula-based methods for mixed-data imputation and synthesis [dissertation]. University of Michigan; 2022.
8. Li X, Zhao R, Fu Y. SYNC: Copula-based synthetic data generation framework. *IEEE Trans Knowl Data Eng.* 2020; 32 (10): 1952–1965.
9. Zhang C, Shi X, Li H, Zhang B. Latent Gaussian copula models for mixed data. *J Mach Learn Res*. 2018; 19 (1): 1–30.
10. Zhang C, Li H, Zhou H. Copula-based modeling for mixed-type biomedical data. *Stat Med*. 2019; 38 (19): 3492–3510.
11. Uddin MA, Debnath M, Roy S, Adiba S, Talukder MMA. Identifying smoking-related predictors on heavy vehicle accidents using logistic regression. *Adv Civ Eng*. 2023; 2023: 7116057. https://doi.org/10.1155/2023/7116057
12. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004; 66 (3): 411–21. https://doi.org/10.1097/00006842-200405000-00021, https://doi.org/10.1097/01.psy.0000127692.23278.a9
13. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, *et al*. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013; 14 (6): 451–460. https://doi.org/10.1038/nrn3502